Predicting Memorability of Images Using Attention-driven Spatial Pooling and Image Semantics

Bora Celikkale[a], Aykut Erdem[a], Erkut Erdem[a,*]

[a]*Department of Computer Engineering, Hacettepe University, Ankara, Turkey*

Abstract

In daily life, humans demonstrate an amazing ability to remember images they see on magazines, commercials, TV, web pages, etc. but automatic prediction of intrinsic memorability of images using computer vision and machine learning techniques has only been investigated very recently. Our goal in this article is to explore the role of visual attention and image semantics in understanding image memorability. In particular, we present an attention-driven spatial pooling strategy and show that considering image features from the salient parts of images improves the results of the previous models. We also investigate different semantic properties of images by carrying out an analysis of a diverse set of recently proposed semantic features which encode meta-level object categories, scene attributes, and invoked feelings. We show that these features which are automatically extracted from images provide memorability predictions as nearly accurate as those derived from human annotations. Moreover, our combined model yields results superior to those of state-of-the art fully automatic models.

Keywords: image understanding, image memorability, visual saliency, spatial pooling, semantic features

*Corresponding author at Department of Computer Engineering, Hacettepe University, Beytepe, Cankaya, Ankara, Turkey, TR-06800. Tel: +90 312 297 7500, 146. Fax: +90 312 297 7502. Email address: erkut@cs.hacettepe.edu.tr.

Email addresses: ibcelikkale@cs.hacettepe.edu.tr (Bora Celikkale), aykut@cs.hacettepe.edu.tr (Aykut Erdem), erkut@cs.hacettepe.edu.tr (Erkut Erdem)

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 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 [1, 2]. We also have an exceptional visual memory [3, 4] that we can remember particular characteristics of a scene with ease even if we look at it only a few seconds [5]. Here, what is being remembered is considered nothing like an identical representation of the scene itself but the gist of it [6, 7]. 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 [8], and the image semantics such as existing objects and their spatial relationships.

Interestingly, we can recall some images surprisingly well while some are lost in our minds. Put simply, not all images are equally memorable. Isola et al. [9] were the first to carry out a computational study about this phenomenon, the so-called intrinsic memorability of images. They devised a Visual Memory Game experiment and utilized Amazon's Mechanical Turk service to quantify the memorability of 2222 natural images (see Figure 1). In the course of these experiments, a total of 665 participants were shown a sequence of images, each of which was displayed for 1 second with a short gap in between image presentations. These subjects were then asked to provide a feedback any time whenever he/she thinks an identical image is displayed. By this setup, a memorability score for each image is calculated by the rate at which the subjects detect a repeated presentation of it. The authors showed that the memorability of an image is pretty consistent across subjects and under a wide range of contexts, which indicates that image memorability is in fact an intrinsic property of images. In addition, the authors explored the use of different visual features and interestingly showed that the intrinsic memorability of an image can indeed be estimated reasonably well by a machine. Since that seminal work, there has been only a few works that explore this difficult and interesting problem [10, 11, 12, 13, 14].

Our first goal in this study is to explore the role of visual attention in

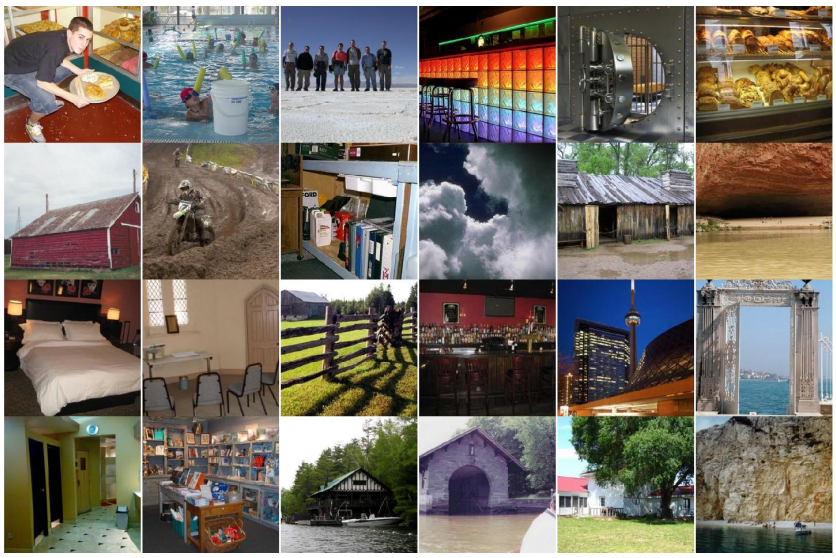


Figure 1: Sample images from the MIT memorability dataset [9]. The images are sorted from more memorable (top left) to less memorable (bottom right).

understanding image memorability. We humans use attentional mechanisms to efficiently perform higher level cognitive tasks by focusing on a small and relevant bits of the visual stimuli. Our intuition is that we are perhaps more likely to remember or forget an image depending on which parts of the image we focus more. To give an example, Figure 2 illustrates the function of visual attention in selecting important features from images. Suppose that we are exposed to these three natural images, each having different visual contents, i.e. different objects, scene characteristics. Our visual system focuses on certain regions that attract our attention as modeled here by a bottom-up saliency model. In this work we propose a visual attention-driven spatial pooling strategy to select important features from images. Our approach makes use of two complementary feature pooling schemes related to visual attention. First, we investigate selecting features from the most salient regions of the images determined according to a recently proposed bottom-up visual saliency model [15]. Our second scheme, on the other hand, considers a top-down definition of visual attention and employs an object-centric spatial pooling scheme. To our interest, a body of research in cognitive sciences promotes that attention plays an important role in understanding natural scenes and enhancing visual memory [7, 16, 17, 18, 19]. However, none of the previously proposed memorability models make use of any attentional

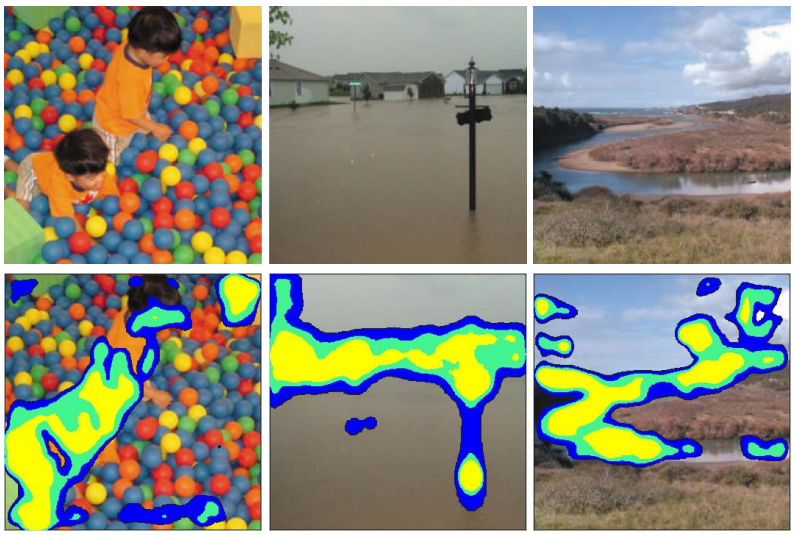


Figure 2: Top: Examples for the most memorable (left), typically memorable (middle), least memorable (right) images in the MIT memorability dataset. Bottom: Salient regions of the images extracted by the method in [15]. The color coding shows the strength of saliency with yellow, green and blue regions corresponding to top 10%, 20%, %30 most salient parts, respectively.

mechanisms for feature selection, and only [11, 13] use saliency maps but as additional image features.

Apart from the global dense image features, some previous studies on image memorability [9, 10, 11, 12] have also investigated the use of high-level semantic information about images. They consider objects-related features [9, 12], presence of certain object and scene categories [9, 10, 11], and their attributes [10], which are all based on manual annotations produced by humans. Figure 3 illustrates some sample images from the MIT memorability dataset along with the semantic features that are manually collected from the human subjects [10]. As illustrated here, an image can be semantically represented in terms of objects, scene information and related attributes.

In addition to our attention-driven feature selection strategy, our second focus in this study is to investigate the use of a diverse set of recently proposed semantic features which encode meta-level object categories [20], scene attributes [21], and invoked feelings [22] for predicting image memorability. Compared to the features considered in the former studies [9, 10, 12], these semantic features can be directly extracted from the images, eliminat-

Object: *person, seat, bottle, chair, floor*

Scene: *indoor, casino, sports and leisure*

Attribute: *has_person, attractive, pleasant, individual, routine, sitting, clear_glasses, ...*

Object: *person, wall, chandelier, ceiling lamp*

Scene: *indoor, shopping and dining, bakery/shop*

Attribute: *has_person, standing, people go, is_interesting, group, routine, ...*

Object: *mountain, sky, tree, natural elevation*

Scene: *outdoor natural, mountains hills desert sky*

Attribute: *peaceful, is_interesting, hang_on_wall, exciting, famous, ...*

Figure 3: Top: Examples for the most memorable (left), typically memorable (middle), least memorable (right) images in the MIT memorability dataset. Bottom: Sample human annotated attributes as collected in [10].

ing the need for manual annotations. Using these features thus decreases the complexity of the prediction process and makes the prediction model to work in a fully automatic manner. Moreover, compared to prior work, these features encode semantic properties of images from a perspective or scale that has not been investigated before. The Meta-class descriptor [20] encodes image semantics based on a hierarchical structure of object categories (concepts) by capturing the relationships among them. The SUN Scene Attributes [21] represents an image by means of responses of a comprehensive list of attribute classifiers that relates to different scene characteristics such as affordances, materials and surface properties. The SentiBank features [22] are the responses of a set of classifiers trained to detect adjective-noun pairs (attributes - objects), and used to associate certain sentiments with images.

In order to validate our approach, we performed a series of experiments on the MIT memorability dataset. To show the effectiveness of the attention-driven pooling strategy, we used the dense global features employed in [9], namely SIFT [23], HOG [24], SSIM [25] and we analyzed the gain when the features pooled over the salient regions are concatenated to the feature vectors obtained with spatial pyramid pooling [26]. Moreover, regarding

our second goal, we performed experiments with the high-level semantic features [20, 21, 22] and tested their performances on predicting image memorability. Lastly, we compared our combined model, which uses both semantic features and dense global features pooled over salient regions and spatial pyramids, to the state-of-the-art models in the literature.

Our main contributions are: (1) an attention-driven pooling approach to put special emphasis on the interesting parts of the images in the computations, (2) a systematic analysis of a diverse set of semantic features on predicting image memorability, and (3) experiments demonstrating that the combination of these ideas provides significant improvement over the existing fully-automatic models.

2. Related Work

All the existing image memorability models in the literature, including our approach, follow the general framework in [9]. In the training step, some low and high-level visual features are extracted from the images and they are used together with the corresponding ground-truth memorability scores to train a support vector regression (SVR) machine, which can then be used to predict the memorability score of a given test image (Figure 4). In [9], the authors suggested representing images by means of some low-level image features such as SIFT [23], HOG [24], SSIM [25], GIST [8] and color histograms, and/or some semantic features which can be extracted from object and scene annotations. Their proposed model predicts image memorability significantly better than chance, illustrating that such image memorability models can be developed. Since then, a number of models [10, 11, 12, 13] have been proposed to improve the results of [9]. In general, these recent studies examine the prediction problem by investigating new features that the authors consider to be relevant to intrinsic memorability of images.

One of our goals in this study is to explore the function of visual attention in predicting intrinsic memorability of images. In that respect, our work shares some motivating factors with the models suggested in [11, 13]. In [11], the authors presented a probabilistic model to measure memorability of image regions, which can be used to predict image memorability as well as the regions that are more likely to be remembered. Within their framework, they suggested to use saliency maps of images as features along with some other visual features. In [13], the authors performed an eye-tracking experiment on a subset of the images in order to observe which parts of those images attract

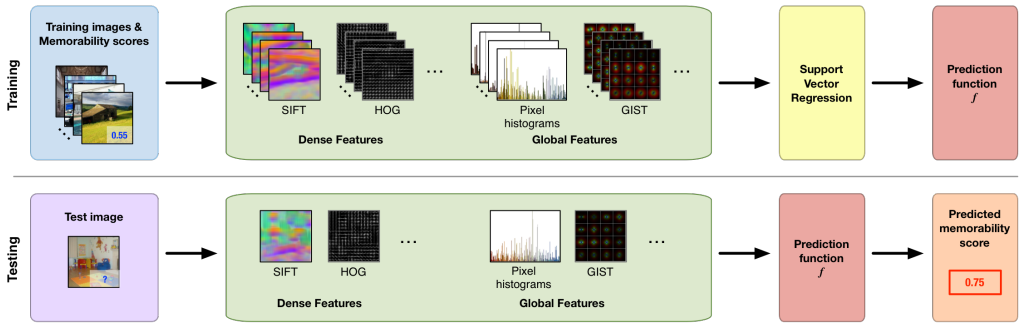


Figure 4: The common training and testing pipeline for learning image memorability.

subjects' attention. They have observed that there is a strong correlation between fixation durations and the memorability scores. In addition, the authors proposed two attention-guided (saliency-oriented) features which are shown to be useful in predicting image memorability.

Beyond visual attention-based features, we specifically aim to investigate the use of attentional mechanisms for selecting relevant features to image memorability. Previous models [11, 13] employ saliency maps or saliency-oriented features as additional images features. In contrast, our key insight is that the visual content in the regions that attract attention is as important as or even more valuable than the whole image content in predicting intrinsic memorability of an image. Thus, whereas prior work [9, 11, 13] employs a fixed pooling layout for feature pooling, we propose to consider a pooling scheme that focuses on salient regions within images. We expect this additional feature selection mechanism will allow us to capture characteristics of images relevant to memorability and accordingly improve the prediction performances of dense image features. The details of our feature pooling scheme will be given in Section 3.

In this work, we also consider ways to boost the success of memorability predictions by employing high-level descriptors that encode the semantic content of images. Similar to [9, 10, 11, 12], we make use of information regarding to objects in images, scene knowledge and/or attributes. In [10], Isola et al. investigated the use of annotated visual attributes to estimate memorability of images. Their study revealed that exploiting available human-describable attributes greatly increases the quality of the predictions. To deeply understand which attribute is a better indicator of memorability, they investigated a greedy feature selection approach to select the best set of relevant

attributes. In another study [12], the authors proposed two novel spatial features which can be extracted from the object annotations exist in the dataset. While the first feature measures the importance of the object in terms of how close it is to the image center and how large it is, the second feature is related to how much unusual the coverage of the object is among all other objects from the same visual category. Their results show that both of these features improve the memorability prediction accuracy.

As compared to [10, 12], however, the semantic features that we employ, which encode meta-level object categories [20], scene attributes [21], and invoked feelings [22], have quite a number of distinct benefits. While the semantic features used in the previous models are based on manual image annotations collected from human subjects, these features can be automatically extracted from the images. This allows us to develop a prediction model which can work in the absence of this sort of high-level annotations. Our approach, thus, requires no supervision and has dramatically less complexity in the training and testing. Notably, among the previous studies, only [11] employed such an automatically extracted semantic feature which is composed of the responses of many pre-trained generic object detectors from ObjectBank [27]. However, these ObjectBank features can be considered as limited as compared to our features, specifically the meta-level object categories [20] which represent an image by means of abstract classes of objects in a hierarchical structure obtained by grouping similar object classes and putting forward higher level common features. The details of our semantic features will be given in Section 4.

To our knowledge, no previous work attempted to improve image memorability prediction based on an attention-guided feature selection mechanism. This article expands upon our previous workshop publication [28]. In this version, we add an entirely new set of experiments on the MIT Memorability dataset, and perform a more thorough experimental analysis to validate that selecting features from the salient image regions via our proposed attention-driven pooling strategy can indeed make more accurate predictions of memorability scores. In addition, we study a group of semantic features related to meta-level object categories [20], scene attributes [21], and invoked feelings [22] that can automatically extracted from images (Section 4), and analyze their roles in predicting memorability of images. Thus, we provide additional discussion of the results and related work, and include new quantitative comparisons of our combined framework against the state-of-the-art.

3. Attention-driven Spatial Pooling

The memorability model by Isola et al. [9] and the follow-up studies [11, 13] all employ spatial pyramid (SP) based pooling [26] for dense global features (Section 3.1). In this study, we propose a complementary visual attention-driven spatial pooling scheme for image memorability, which allows us to select features from the salient image regions. In particular, these regions are estimated by considering two different saliency maps. While one of them is estimated via a bottom-up saliency model (Section 3.2), the second one is derived from a complementary object-level saliency map which captures information about foreground objects in the images (Section 3.3). The details of our proposed attention-driven pooling strategy are given in Section 3.4.

3.1. Spatial Pooling

The common pipeline of modern visual recognition tasks uses spatial pooling in order to construct compact representations and achieve robustness to noise and clutter. After extracting local or global low level features from images, feature vectors are encoded to codewords using a descriptive vocabulary. Then, histograms of these codewords are computed in order to get the fixed-length exemplar vectors of the predefined subregions of the image. Final representation is formed by simple concatenation of all histogram vectors obtained in this way. Boureau et al. [29] showed various factors that affect the performance of pooling strategies and demonstrated the importance of the step. For example, Isola et al. [9] used simple 2-level spatial pyramid pooling strategy in their work. However, in this study, we approach the pooling step by further incorporating visual attention mechanisms with the inspiration that visual attention is considered highly related with memorability [7, 16, 17, 18, 19].

3.2. Visual Saliency

In recent years, there has been an increasing interest in computational models of visual saliency estimation and their use for several computer vision tasks. Starting from the seminal work by Itti, Koch, and Niebur [30], most of the existing models consider a bottom-up strategy. First, center-surround differences of various features at multiple scales are computed for each feature channel. Then the final saliency map is formed by linearly combining feature maps after a normalization step. For a recent survey, please refer

to [31]. In our experiments, we employed the publicly available implementation of a recently proposed saliency model [15][1], which examines the first and second-order statistics of simple visual features such as color, edge and spatial information.

Consider Figure 5(a) where we present the result of the bottom-up saliency estimation for a sample image. From the saliency map 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 the strongly correlated with the interestingness of the regions [32, 33] in the sense that while the most salient patches captures the children, the least salient ones mostly correspond to background or those regions which have little importance in terms of image content.

3.3. Objectness Measure

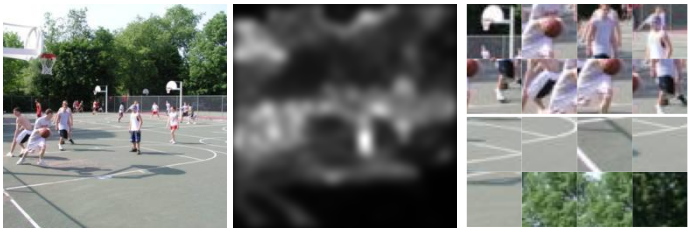
In [34], Alexe et al. introduced a generic (category-independent) objectness measure[2] to quantify how likely an image window contains an object. In more detail, the authors first analyzed several image cues, namely multi-scale saliency, color contrast, edge density (near window borders) and superpixel straddling, each of which were shown to be an indicator of objectness, but to a certain degree. Then they proposed a 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.

Figure 5(b) shows some sample patches sampled from the object-level saliency map as we did for the bottom-up saliency. Similarly, the rightmost top two rows of patches taken from salient regions mostly correspond to the mill in the image, which is the most salient object. Other non-salient patches correspond to unimportant areas such as the sky or the field.

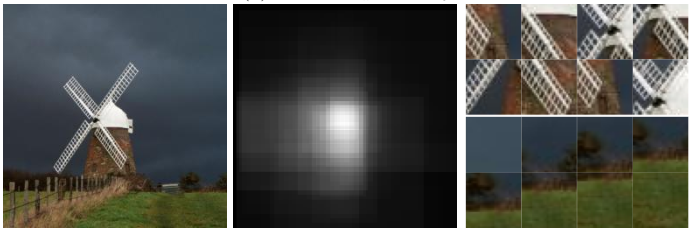
---

[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/`



(a) Bottom-up saliency



(b) Object-level saliency

Figure 5: 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.

3.4. Proposed Strategy

Instead of using a fixed pooling layout like the spatial pyramid structure used in [9], we propose an image-driven pooling strategy by considering salient regions of the images. For this purpose, we both utilize the bottom-up and object-level saliency maps described in the previous subsections. In this way, our pooling method adaptively focuses solely on the image regions that attract attention, ignoring not important, non-attractive parts of the images.

The system diagram of the proposed pooling approach is given in Figure 6. First, dense visual features are extracted from the input image. Low level dense 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 regions possibly containing important foreground objects. Next, to form histogram-based visual descriptors, the encoded vectors are pooled together over the extracted attention-driven spatial layouts.

For the prediction pipeline for spatial pooling, we used the following steps:

(1) Feature Extraction. For an image **I**, we obtain a global descrip-

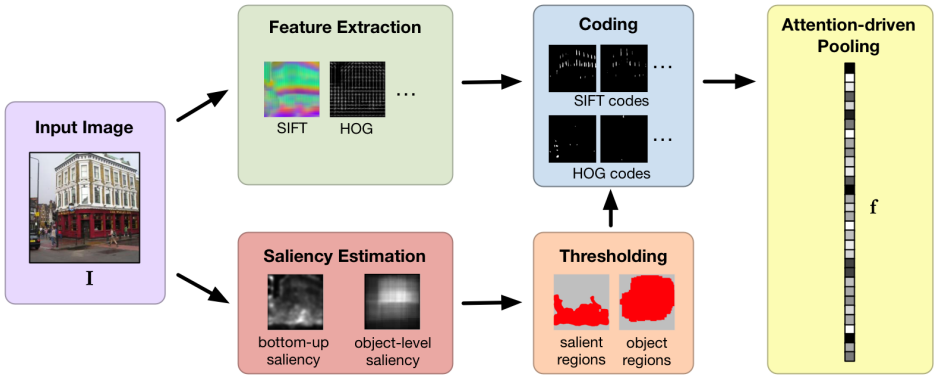


Figure 6: The proposed visual attention-driven spatial pooling pipeline for image memorability.

tion of $\mathbf{I}$ by extracting D-dimensional local features such as SIFT [23], HOG [24], SSIM [25] at N different locations, denoted with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$. 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.

(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 $\mathbf{c}_i = [c_1^i, c_2^i, \ldots, c_K^i]^T$ by applying vector quantization. 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 [15], which was shown to provide state-of-the-art performance in predicting eye fixations. For the object-level saliency map, we randomly sample many windows from $\mathbf{I}$ and measure the objectness of these image windows by using the generic objectness measure proposed in [34]. Then we compute an objectness score for each pixel by averaging over all the scores of the windows which contain that pixel to obtain the generic objectness map of $\mathbf{I}$.

(4) Pooling. In the pooling step, instead of considering a fixed image-independent set of spatial regions, as employed in [9], here we propose to use image-specific spatial regions for feature pooling. Specifically, we locate the regions of interest 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 7 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 $\mathcal{R}$:

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

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. Semantic Features

As discussed in Section 2, [9] showed that memorability of an image is highly correlated with the semantic content of the image. For instance, only making use of manual annotation of object and scene labels is shown to give pretty good results. In a follow-up work [10], the authors collected attributes that humans used to describe images and explored their role in determining the intrinsic memorability of images. Motivated from these findings, here, our goal is to extend our framework to include automatically extracted semantic attributes. For that purpose, we propose to use three recently proposed semantic descriptors: The Meta-class descriptor [20] provides object-specific high-level information about image content (Section 4.1). The SUN Scene Attributes [21], on the other hand, characterize the images by means of a set of hand-picked functional, material, surface and spatial properties (Section 4.2). Finally, the SentiBank features [22] are used to include feelings that are invoked in a viewer into the computations (Section 4.3).

Figure 7: Visual attention-driven feature pooling scheme. For a given image, a bottom-up saliency map and an object-level saliency map are estimated and then the feature vectors are pooled over the salient regions of the images (depicted as bright areas in the images).

4.1. Meta-class Features

In computer vision, attributes typically denote properties that humans use to verbally describe the visual content such as individual objects, object classes, scenes. Besides, they can also indicate properties shared among different object classes. The Meta-class descriptor [20] falls under this category that it captures common visual properties of different sets of object classes and represents an image in terms of them. In essence, these abstract categories are linear combinations of multiple non-linear classifiers trained on different low-level features. The authors trained a tree of classifiers using a subset of ImageNet [35] dataset and with the help of predefined object classes from ILSVRC2010 and Caltech256 datasets. Each node in the tree correspond to a meta-class obtained by combining two previously determined meta-classes (i.e. a set of object classes) which makes them easy to distin-

guish from other sets of object classes. They demonstrate that this descriptor gives state-of-the-art results for object categorization against similar semantic representations such as Object-Bank [27] and PiCoDes [36].

In our work, we use Meta-class features, i.e. the responses of the learned tree of classifiers, to obtain a semantic representation of image content by means of the presence or absence of the meta-classes. Figure 8 demonstrates the importance of certain object classes in determining the memorability of an image on some sample images from the MIT memorability dataset. It can be easily observed that the most memorable images generally are those that contain close-up human faces. Interestingly, typical memorable images generally do have humans and/or human-made structures or objects at a distance. The least memorable images are mostly the images of natural scenes.

4.2. SUN Scene Attributes

In [21], Patterson and Hays carried out a large scale experiment to form a scene attribute dataset by crowdsourcing. They collected 102 discriminative attributes related to different visual properties of a scene, namely affordance, material, surface and spatial envelope properties. Using these collected attributes as ground truth, they also trained a binary classifier for each attribute and proposed to use responses of these classifiers to obtain an attribute based scene representation. They showed that this intermediate level representation captures scene content information remarkably well and can be effectively used for different computer vision tasks including scene classification, automatic image captioning, semantic image retrieval.

In our framework, we use the confidence scores of the scene attribute classifiers as complementary semantic features for learning image memorability. Figure 9 illustrates some of the most confident scene attributes [21] that are extracted from some sample images having different memorability scores. We observe that the most memorable images are typically associated with the "*no-horizon*", "*enclosed-area*", "*cloth*" and "*man-made*" attributes whereas the least memorable ones mainly have "*open-area*", "*grass*", "*vegetation*" and "*natural*" attributes. These observations are in accordance with the findings reported in [9, 10] suggesting that the images of people and enclosed spaces are more memorable images than those of natural images.

Figure 8: Sample images from memorability database. Top row shows some samples from the most memorable images which mostly contain close-up human faces by large. Middle row shows samples from typically memorable images which generally have humans and/or human-made structures or objects at a distance. Bottom row shows the least memorable samples which are mainly the images of natural scenes.

4.3. SentiBank

Borth et al. [22] recently proposed a large scale visual sentiment ontology based on the psychological theory of Plutchick's Wheel of Emotions [37]. To construct this ontology, the authors followed a data-driven approach and used a large set of tagged images and videos from the web to gather a list of adjectives and nouns based on their co-occurrences with each of the 24 emotions defined in [37]. They assigned certain sentiment values to these tags and employed them to form Adjective-Noun Pairs (ANPs) which reflect strong emotions and frequently appear together. Then, they trained a classifier for each ANP using some low and high-level visual features. They finally selected 1200 of those trained ANP classifiers that have a reasonable

cloth, enclosed area, no-horizon, electric/indoor lighting

man-made, no-horizon, enclosed area, natural light

far-away horizon, natural light, open area, grass, natural, vegetation

Figure 9: Sample images from memorability database for most memorable (left), typically memorable (middle) and least memorable (right) with their most confident scene attributes predicted by [21].

classification performance to build their visual sentiment analysis framework known as the SentiBank.

In our approach, we use the visual sentiment classifiers from the SentiBank to include emotion-based semantic features to our image representations. Figure 10 demonstrates some sample images with different memorability scores with the associated ANPs as predicted by the SentiBank classifiers. As can be observed, in each case, the classifiers accurately capture the feelings invoked in the viewers. Although there is no common pattern for ANPs associated with images from different memorability levels, we observe that in general, the most memorable images are linked with the emotions that can relate to humans (e.g., *shy smile*). Moreover, the typically memorable images invoke feelings related to man-made structures (e.g., *calm pond*) whereas the least memorable ones are associated with ANPs related to natural scenes (e.g., *beautiful garden*).

5. Experiments

In this section, we first give brief details about our experimental setup and then demonstrate the effectiveness of the proposed approach through a series of experiments.

5.1. Experimental Setup

For the quantitative analysis we used Spearman's rank correlation measure (ρ). The performance was evaluated over 25 different splits of the dataset

shy smile, traditional dress, innocent smile | *warm creek, smooth water, calm pond* | *beautiful garden, sunny trees, pretty scene*

Figure 10: Sample images from memorability database for most memorable (left), typically memorable (middle) and least memorable (right) with their most confident sentiment ANPs as predicted by [22].

containing 1111 training and 1111 testing images (the same splits used in [9]). 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's rank correlation to this score. In addition, the performance of a model can be analyzed at different memorability levels by ranking the test images according to their predicted memorability scores and then computing the cumulative average of empirical memorability scores at different quantiles. For instance, a good image memorability model should have cumulative averages close to 1 for the top most memorable images predicted by a model and close to 0 for the bottom least memorable images.

5.2. Results and Discussions

In the first part of the experiments, we analyzed the performance of our proposed attention-driven pooling scheme in detail. We conducted our experiments on three global dense features, SIFT [23], HOG [24] and SSIM [25], which were used in [9]. Specifically, we analyzed the performance when features obtained with our attention-driven pooling strategy are concatenated to those derived by the standard spatial pyramid pooling. We examined the prediction accuracy of each dense feature separately. We also provided the results for the combination of these features. We separately trained different SVRs to map from the features pooled over these maps to memorability scores.

A summary of our results is given in Table 1. As can be seen, the attention-driven pooling alone performs poorly as compared to the 1-level

Table 1: Comparison of pooling schemes (1-level Spatial Pyramid pooling (SP), Attention-based Pooling (AP) and their combination) using dense global features SIFT, HOG and SSIM. Results are given as the average empirical memorability scores reported for the top 20, top 100 highest and bottom 20, bottom 100 lowest predicted memorability scores and the Spearman's Rank Correlation (ρ) values.

		SIFT	HOG	SSIM	SIFT+HOG+SSIM
SP	Top 20	83.8%	83.3%	83.2%	85.0%
	Top 100	82.3%	81.9%	80.7%	80.5%
	Bottom 100	54.9%	56.0%	56.7%	54.6%
	Bottom 20	50.3%	47.9%	54.0%	50.1%
	ρ	0.430	0.431	0.436	0.458
AP	Top 20	87.6%	87.8%	84.9%	87.4%
	Top 100	81.8%	83.0%	83.4%	83.7%
	Bottom 100	56.6%	55.9%	56.7%	55.6%
	Bottom 20	58.2%	48.4%	56.4%	51.8%
	ρ	0.390	0.420	0.427	0.438
SP + AP	Top 20	86.0%	86.9%	86.8%	86.9%
	Top 100	83.3%	82.9%	81.0%	82.6%
	Bottom 100	55.7%	54.8%	53.6%	53.4%
	Bottom 20	49.9%	47.4%	48.5%	53.2%
	ρ	0.435	0.448	0.454	0.472

spatial pyramid (SP) based pooling. However, for each dense feature, there is a notable improvement in the performance with the inclusion of our attention-driven pooling scheme to the SP based baseline. More specifically, the SSIM feature has the most significant gain where the correlation moves from $\rho = 0.436$ to $\rho = 0.454$. Furthermore, we observed that the result of the combined features can be also improved when our pooling strategy is used. However, the amount of gain, from $\rho = 0.458$ to $\rho = 0.472$, is relatively smaller than those of single features. When the average memorability scores of the models are examined at top 20/100 and bottom 20/100 quantiles, we have similar observations. In conclusion, the combined pooling framework performs especially much better by assigning less memorable images lower scores. These results support our claim that the image regions which retain in human memory are correlated with the areas that attract our attention.

In our second experiment, we included the semantic features, namely the

Meta-class features [20], the SUN scene attributes [21] and the SentiBank features [22] to the original feature pool (pixels, GIST, SIFT, SSIM, HOG-based image features), and performed a thorough analysis of the framework with all possible combinations of these features and pooling strategies.

Table 2 demonstrates the results obtained by SSIM (best performing low-level image feature), our semantic features and their combination. One key observation is that the Meta-class features and the Scene Attributes provide fairly good predictions as compared to the SentiBank or any other low-level cues. In particular, the Meta-class descriptor alone achieves approximately $\rho = 0.49$ correlation value, which shows us that memorability of images are not only related to single object properties but also to inherent and shared characteristics of different object classes. Similarly, the Scene Attributes alone give nearly $\rho = 0.48$, illustrating the importance of scene properties over objects in the images for image memorability. We achieved the best performance when we combined all semantic features and SSIM with a combination of our proposed attention-driven pooling and 2-level spatial pyramid pooling for the dense features. With this model of ours, the Spearman's rank correlation between the ground-truth ranking and the predictions is estimated as $\rho = 0.515$. This correlation value is smaller than the correlation among humans ($\rho = 0.75$) but it is the best result reported in the literature so far by a fully automatic scheme that does not use any manual object, scene or attribute annotations. It also demonstrates the importance of high-level semantic features as incorporating them increases the rank correlation score from $\rho = 0.472$ (SP+AP) to $\rho = 0.515$ (All). Moreover, the increases in the top 20 and top 100 average memorability predictions support the hypothesis that the semantic content of images is highly correlated with their intrinsic memorability.

In Table 3, we compare the result of our proposed method with the methods of Isola et al. [9], Khosla et al. [11] and Mancas and Le Meur [13]. Our method has the best performance among these state-of-the-art fully automatic approaches. While Khosla et al. [11] achieved $\rho = 0.50$ with their global model which additionally considers memorability characteristics of the local image regions, our model achieves slightly better results with far less complexity. Moreover, another key observation from Table 3 is that most of the memorability prediction schemes predict top memorable images with high precision. For the top 20 and top 100 images, the models have obtained nearly the same average empirical memorability values, which are very close to the scores of human subjects. However, predicting whether an image is less

Table 2: Comparison of the best local dense feature (SSIM) and all semantic features. Results are given as the average empirical memorability scores reported for the top 20, top 100 highest and bottom 20, bottom 100 lowest predicted memorability scores and the Spearman's Rank Correlation (ρ) values.

	SSIM	Scene Attributes	Meta-class	SentiBank	All
Top 20	86.8%	86.4%	86.8%	85.7%	85.0%
Top 100	81.0%	83.7%	81.5%	82.5%	83.3%
Bottom 100	53.6%	54.2%	53.3%	54.8%	52.2%
Bottom 20	48.5%	51.3%	46.7%	47.1%	47.4%
ρ	0.454	0.477	0.487	0.449	0.515

Table 3: The first four rows indicate average empirical memorability scores over different memorability levels. (ρ) is the Spearman's rank correlation between predictions of existing fully automatic models and the empirical results. The best automatic prediction result is shown in bold.

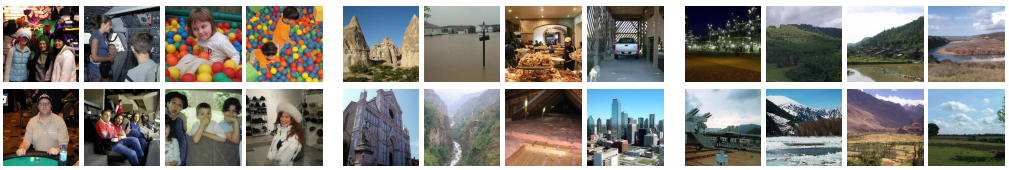
	Isola et al. [9]	Khosla global [11]	Khosla local+global [11]	Mancas & Le Meur [13]	Our Approach	Human subjects
Top 20	83%	84%	85%	–	85%	86%
Top 100	80%	80%	81%	–	83%	84%
Bottom 100	56%	56%	55%	–	52%	47%
Bottom 20	54%	53%	52%	–	47%	40%
ρ	0.46	0.48	0.50	0.48	**0.52**	0.75

memorable is a more difficult problem. In that respect, our model provides better predictions for the bottom 20 and bottom 100 images as compared to the state-of-the-art models. It is important to note that there is still a large gap between our result and that of human subjects in predicting the less memorable images.

Finally, in order to demonstrate the effectiveness of our proposed combined model, we compare our result with those of the human annotations reported in [14]. For object semantics, the authors in [14] achieved $\rho = 0.47$ whereas we obtained a correlation value of $\rho = 0.49$ with the Meta-class descriptor that describes abstract object classes. This shows that fully automatic approaches can also capture object semantics to some extent to improve memorability predictions. On the other hand, the model based on the attribute annotations, gives a better correlation value of $\rho = 0.52$ as compared to those of SUN Scene Attributes and SentiBank features respectively

corresponding to $\rho = 0.48$ and $\rho = 0.45$. Moreover, the model which considers the combined overall semantics (objects + scenes + attributes) has a correlation value of $\rho = 0.54$, which is higher than that of our proposed combined model having $\rho = 0.52$. However, we observe that our model provides better predictions especially for the least memorable images. For the bottom 100 and 20 images while the average ground-truth memorability scores are %55, and %51 for object, scene and attribute annotations, respectively, ours are %52, and %47 which are much closer to the human subjects. Overall, human annotations still have advantage over automatic attributes, however the gap is small. Considering the cost of gathering annotations from human subjects, our approach gives similar performance with much less computational effort.

In Figure 11, we additionally show sample images for different memorability levels predicted by the proposed framework. Figure 12 shows some images on which the memorability predictions based on our approach are incorrect as compared to the empirical results. The reasons for this could lie in the inaccurate predictions of the semantic content or focused regions of images. In Figure 13(a)-(b), for example, we provide the bottom-up and object-level saliency maps of two of the images from Figure 12 together with their memorability maps as computed by the protocol used in [9, 11]. In the memorability maps, the red regions illustrate the objects that contribute positively to the predicted memorability of the image and the blue regions show the objects that contribute negatively to the predicted memorability. In an ideal case, the predicted salient regions need to correspond to the image regions that affect the memorability scores positively or negatively. For the image in Figure 13(a) whose memorability rank is overshot by the proposed prediction scheme, our pooling method can not correctly identify the object regions that correlate with the image memorability. For the image in Figure 13(b) whose memorability rank is undershot by the proposed scheme we observe a similar behaviour for the detection of important object regions. Our pooling scheme gives prominence to only some parts of the object regions that affects the memorability predictions negatively. These imperfect predictions of the important image regions make the features collected via our attention driven pooling scheme cover the image content in an inaccurate way, affecting the estimated memorability scores.



Predicted most memorable (85%) Predicted typical memorable (68%) Predicted least memorable (40%)

Figure 11: Memorability predictions by the proposed 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.

6. Conclusion

Predicting whether an image will be remembered or not is a challenging problem for computer vision. In this paper, we describe our efforts to develop a new fully automatic model for estimating image memorability, which benefits from a novel feature pooling strategy based on visual attention and a set of semantic features that encode meta-level object categories, scene attributes, and invoked feelings.

Our proposed feature pooling strategy is derived from the observation that main memorable areas of an image are the ones that attract the most attention. Different from the fixed pyramidal structure as in [9, 11, 13], our regression model learns memorability scores of images by additionally taking into account the features pooled over the saliency maps. In our pooling scheme, we employed two saliency maps, one obtained by a bottom-up saliency model [15] and the other by a generic objectness model [34], respectively modeling bottom-up and top-down attentional influences on image memorability. Our experiments demonstrated that for the global dense features the combination of classical SP based pooling with the proposed pooling scheme improves the prediction quality.

Moreover, we investigate the use of three recently proposed semantic features, namely the Meta-class [20], the SUN Scene Attributes [21] and SentiBank [22] features, all of which can be automatically extracted from the images. These high-level features are used to describe the presence of certain abstract object categories, attributes related to functional, material, surface properties of scenes, and the emotions induced by the images as captured by the specific adjective-noun pairs. The inclusion of these semantic features

Predicted too high (+793/1111)

Predicted too low (-873/1111)

Figure 12: 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.

into the computations greatly improves the prediction performance that we obtained superior results on the MIT Memorability dataset than those of the fully automatic the state-of-art models.

For highly memorable images, the existing approaches to predict image memorability can yield estimates close to the ground-truth scores from human subjects. However, their performances on determining whether an image is unmemorable is currently far from empirical scores. Even though our model provide the best results reported in the literature for predicting the memorability of less memorable images, it is not as accurate as desired. This

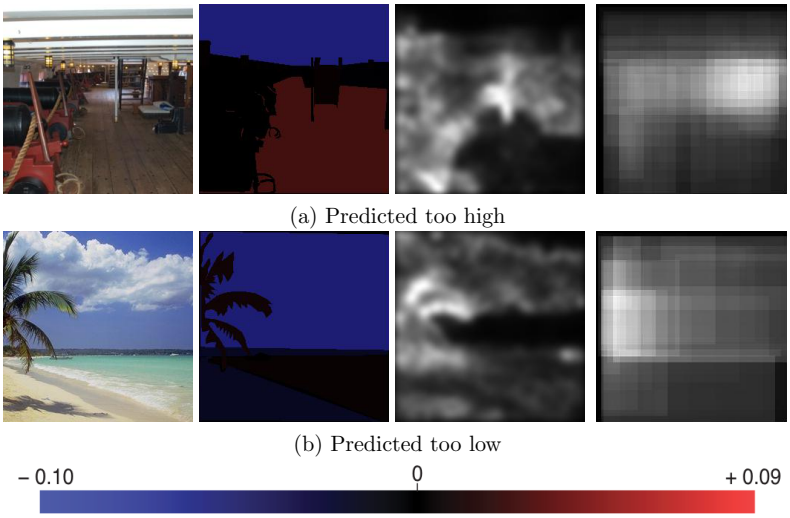


Figure 13: Memorability maps versus bottom-up saliency and object-level saliency maps of two of the images from Figure 12.

opens up possibilities to design or learn new types of features to especially understand less memorable images.

For future work, it would be interesting to investigate how image memorability affects visual saliency. A recent trend in visual saliency estimation is to pose saliency estimation as a supervised learning problem [38, 39, 40, 41, 42]. These models, except [38, 41, 42], 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. In that regard, it would also be interesting to study how to employ semantic attributes for learning saliency, as recently explored in [43] in the context of free-viewing.

Acknowledgments

This work was supported by a grant from The Scientific and Technological Research Council of Turkey (TUBITAK) – Career Development Award 112E146.

References

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

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

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

[4] T. F. Brady, T. Konkle, G. A. Alvarez, A. Oliva, Visual long-term memory has a massive storage capacity for object details, Proc. Natl. Acad. Sci. U.S.A. 105 (38) (2008) 14325–14329.

[5] L. Standing, Learning 10,000 pictures, Quarterly Journal of Experimental Psychology 25 (1973) 207–222.

[6] J. M. Wolfe, Visual memory: What do you know about what you saw?, Current Biology 8 (1998) 303–304.

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

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

[9] P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable?, in: CVPR, 2011, pp. 145–152.

[10] P. Isola, D. Parikh, A. Torralba, A. Oliva, Understanding the intrinsic memorability of images, in: NIPS, 2011, pp. 2429–2437.

[11] A. Khosla, J. Xiao, A. Torralba, A. Oliva, Memorability of image regions, in: NIPS, 2012, pp. 305–313.

[12] J. Kim, S. Yoon, V. Pavlovic, Relative spatial features for image memorability, in: ACM MM, 2013, pp. 761–764.

[13] M. Mancas, O. L. Meur, Memorability of natural scenes: The role of attention, in: ICIP, 2013, pp. 196–200.

[14] P. Isola, J. Xiao, D. Parikh, A. Torralba, A. Oliva, What makes a photograph memorable?, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1469–1482.

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

[16] A. Hollingworth, C. C. Williams, 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) (2001) 761–768.

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

[18] M. A. Cohen, G. A. Alvarez, K. Nakayama, Natural-scene perception requires attention, Psychological Science 22 (2011) 1165–1172.

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

[20] A. Bergamo, L. Torresani, Meta-class features for large-scale object categorization on a budget, in: CVPR, 2012, pp. 3085–3092.

[21] G. Patterson, C. Xu, H. Su, J. Hays, The SUN attribute database: Beyond categories for deeper scene understanding, Int. J. Computer Vision 108 (1–2) (2014) 59–81.

[22] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: Proceedings of the 21st ACM international conference on Multimedia, ACM, 2013, pp. 223–232.

[23] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Computer Vision 60 (4) (2004) 91–110.

[24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, Vol. 1, 2005, pp. 886–893.

[25] E. Shechtman, M. Irani, Matching local self-similarities across images and videos., in: CVPR, 2007, pp. 1–8.

[26] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: CVPR, Vol. 2, 2006, pp. 2169–2178.

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

[28] B. Celikkale, A. Erdem, E. Erdem, Visual attention-driven spatial pooling for image memorability, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Computer Society Conference on, IEEE, 2013, pp. 1–8.

[29] Y.-L. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in: ICML, 2010, pp. 111–118.

[30] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[31] A. Borji, State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 185–207.

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

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

[34] B. Alexe, T. Deselares, V. Ferrari, What is an object?, in: CVPR, 2010, pp. 73–80.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.

[36] A. Bergamo, L. Torresani, A. Fitzgibbon, PiCoDes: Learning a compact code for novel-category recognition, in: NIPS, 2010, pp. 2088–2096.

[37] R. Plutchik, Emotion: A Psychoevolutionary Synthesis, Harper & Row, Publishers, 1980.

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

[39] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: ICCV, 2009, pp. 2106–2113.

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

[41] J. Yang, M.-H. Yang, Top-down visual saliency via joint CRF and dictionary learning, in: CVPR, 2012, pp. 2296–2303.

[42] A. Kocak, K. Cizmeciler, A. Erdem, E. Erdem, Top down saliency estimation via superpixel-based discriminative dictionaries, in: BMVC, 2014.

[43] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, Q. Zhao, Predicting human gaze beyond pixels, Journal of Vision 14 (1) (2014) 1–20.