What makes an object memorable?

Rachit Dubey*[1], Joshua Peterson*[2], Aditya Khosla[3], Ming-Hsuan Yang[4], and Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology [2]University of California, Berkeley [3]Massachusetts Institute of Technology
[4]University of California, Merced

Abstract

Recent studies on image memorability have shed light on what distinguishes the memorability of different images and the intrinsic and extrinsic properties that make those images memorable. However, a clear understanding of the memorability of specific objects inside an image remains elusive. In this paper, we provide the first attempt to answer the question: what exactly is remembered about an image? We augment both the images and object segmentations from the PASCAL-S dataset with ground truth memorability scores and shed light on the various factors and properties that make an object memorable (or forgettable) to humans. We analyze various visual factors that may influence object memorability (e.g. color, visual saliency, and object categories). We also study the correlation between object and image memorability and find that image memorability is greatly affected by the memorability of its most memorable object. Lastly, we explore the effectiveness of deep learning and other computational approaches in predicting object memorability in images. Our efforts offer a deeper understanding of memorability in general thereby opening up avenues for a wide variety of applications.

1. Introduction

Consider the left image in Figure 1. Even though the person on the right is comparable in size to the person on the left, he is remembered far less by human subjects, indicated by their respective memorability scores of 0.18 and 0.64. Moreover, people tend to remember the person on the left and the fish in the center, even after 3 minutes and more than 70 additional visual stimuli have passed. Interestingly, despite vibrant colors and considerable size, the boat is far less memorable with a memorability score of 0.18.

One of the primary goals of computer vision is to aid human-relevant tasks, such as object recognition, object detection, and scene understanding. Much of the algorithms in service of this goal have to make inferences about all objects in a scene. In comparison, humans are incredibly selective in the information they consider from the possible visual candidates they encounter, and as a result, many human tasks are dependent on this filtering mechanism to be performed effectively. For this reason, it is important for vision systems to have information on hand concerning what objects humans deem important in the world, or in our specific case, which of them are worth remembering. Such information holds exciting promise. For example, it can help in building assistive devices (goggles) so that the elderly can easily memorize objects that they tend to forget, or help design better instructional diagrams involving memorable graphic objects.

Figure 1: **Not all objects are equally remembered.** Image showing objects and their respective memorability scores (left) obtained from our experiment. We note that certain objects (the fish and left person) are more memorable than other objects. Right panel shows the ground truth map generated from the object segments and memorability scores.

Going back to Figure 1, why are the fish and left person more memorable and how do these objects influence the overall memorability of the photo? The community has made great strides in understanding comparable visual properties of the world such as saliency [19, 15, 16, 7, 4, 11, 12] and importance [3], but we still do not have a clear understanding of what objects are worth remembering in the world. Although recent studies related to image memorability [17, 22, 24, 18, 8] have explored this at the image-level, no work has explored what exactly in an image is remembered. Using object annotations and predictive models, such knowledge can be potentially inferred from the memorability score of an image alone [25], but

* denotes equal contribution

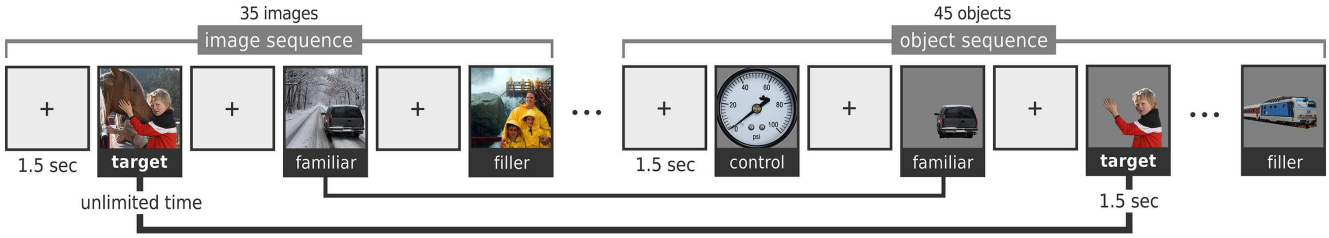

Figure 2: **Object Memory Game.** Participants viewed a series of images followed by a sequence of objects and were asked to indicate whether each object was seen in the earlier sequence of full images.

these methods will ultimately require ground truth object memorability data to be properly evaluated and analyzed. To enable the development of such approaches, we collect ground truth object-level memorability scores and conduct an extensive empirical investigation of memorability at the object level. This allows for a simple yet powerful strategy that provides detailed answers to many interesting questions at hand. While image memorability studies have provided invaluable knowledge, the study of object memorability will enable unique applications in the field of computer vision and computational photography not possible from the study of image memorability alone. It can guide cameras to automatically focus on memorable objects and in the process help take photographs that are more memorable. Similarly, it can enable intelligent ad placement software that embeds products (objects) in adverts in such a way that humans are likely not to forget.

In this paper, we systematically explore the memorability of objects within individual images and shed light on the various factors that drive object memorability. In exploring the connection between object memorability, saliency, object categories, and image memorability, our paper makes several important contributions.

Contributions. **(1)** This paper presents the first work that studies the problem of object memorability and provides a deeper understanding of what makes objects in an image memorable or forgettable. While previous work has tried to infer such knowledge computationally [25], our work is the first to directly quantify and study what objects in an image humans actually remember. **(2)** We uncover the relationship between visual saliency and object memorability and demonstrate those instances where visual saliency directly predicts object memorability and when/why it fails to do so. While there have been a few very recent studies that explore the connection between image memorability and visual saliency [8, 34, 26], our work is the first to explore the connection between object-level memorability and visual saliency. **(3)** We make significant headway in disambiguating the link between image and object memorability. We show that in many cases, the memorability of an image is primarily driven by the memorability of its most memorable object. Furthermore, we show that our compiled dataset can serve as a benchmark for evaluating automated object memorability algorithms and enable/encourage future work in this exciting line of research.

2. Measuring Object Memorability

As a first step towards understanding memorability of objects in images, we compile an image dataset containing a variety of objects from a diverse range of categories. We then measure the probability that every object in each image will be remembered by a large group of subjects after a single viewing. This helps provide ground truth memorability scores for objects inside images (defined as image segments) and allows for a precise analysis of the memorable elements within an image.

Toward this, we utilized the PASCAL-S dataset [30], a fully segmented dataset built on the validation set of the PASCAL VOC 2010 [13] segmentation challenge. To improve segmentation quality, we manually refined the segmentations from this dataset. We removed all homogenous non-object or background segments (e.g. ground, grass, floor, and sky), along with imperceptible object fragments and excessively blurred regions. All remaining object segmentations were tested for memorability. In summary, our final dataset comprises 850 images and $3,412$ object segmentations (i.e. an average of 4 objects per image), for which we gathered ground truth memorability through crowd sourcing.

2.1. Object Memory Game

To measure the memorability of individual objects in each dataset image, we created an alternate version of the Visual Memory Game through Amazon Mechanical Turk following the basic design in [18], with the exception of a few key differences (refer to Figure 2). In our game, participants first viewed a sequence of 35 images one at a time, with a 1.5 second interval between image presentations. The subjects were asked to remember the contents and objects inside these images to the best of their ability. To ensure that subjects would not only just look at the salient or center objects, they were given unlimited time to freely view the images. Once they were done viewing an image, they could press any key to advance to the next image. After the initial image sequence, participants viewed a sequence of 45 objects, their task then being to indicate

through a key press which of those objects was present in one of the previously shown images. Each object was displayed for 1.5 seconds, with a 1.5 second gap between each object in the sequence. Pairs of corresponding image and object sequences were broken up into 10 blocks. Each block consisted of 80 total stimuli (35 images and 45 objects), and lasted approximately 3 minutes. At the end of each block, the subject could take a short break. Overall, the experiment takes approximately 30 minutes to complete.

Unknown to the subjects, each sequence of images inside each block was pseudo-randomly generated to consist of 3 "target" images taken from the PASCAL-S dataset, whose objects were later presented to the participants for identification. The remaining images in the sequence consisted of 16 "filler" images and 16 "familiar" images. Filler images were randomly selected from the DUT-OMRON dataset [39], while the familiar ones were randomly sampled from the MSRA dataset [32]. In a similar fashion, the object sequence in each block was also generated pseudo-randomly to consist of 3 target objects (1 object taken randomly from each previously shown target image). The remaining objects in the sequence consisted of 10 control, 16 filler, and 16 familiar objects. Filler objects were sampled randomly from the 80 different object categories in the Microsoft COCO dataset [31], while the familiar objects were sampled from objects taken from the previously displayed familiar images in the image sequence. The familiars ensured that the subject were always engaged in the task and the fillers helped provide spacing between the target images and target objects. While the fillers and familiars (both images and objects) were taken from datasets of real world scenes and objects, the control objects were artificial stimuli randomly sampled from the dataset proposed in [6]. Control objects were meant to be easy to remember and served as a criteria to ensure quality [6, 18]. Target images and their corresponding target objects were spaced $70 - 79$ stimuli apart, while familiar images and their objects were spaced $1 - 79$ stimuli apart.

All images and objects appeared only once, and each subject was tested on only one object from each target image. Objects were centered within the image they originated from and non-object pixels were set to grey. Participants were required to complete the entire task, which included 10 blocks ($\sim$30 minutes) and could not participate in the experiment a second time. The maximum time that subjects could take to finish the experiment was 1 hour. After collecting the data, we assigned a memorability score to each target object in our dataset, defined as the percentage of correct detections by subjects (refer to Figure 1 for an example). Strict criteria was undertaken to screen subjects' performance and to ensure that our final dataset consisted of quality subjects. We discarded all subjects whose accuracy on the control objects was below 70%. The accuracy of these subjects on filler objects and familiar objects was greater than chance ($> 75\%$) demonstrating that our data consists of subjects who were paying attention to the task at hand. The mean time taken by the subjects to view an image was 2.2 seconds with a standard deviation of 1.6 seconds. In total, we had $1,823$ workers from Mechanical Turk each having at least 95% approval rating in Amazon's system. On average, each object was scored by 16 subjects and the average memorability score was 0.33 with a standard deviation of 0.28.

2.2. Consistency Analysis

To assess human consistency in remembering objects, we repeatedly divided our entire subject pool into two equal halves and quantified the degree to which memorability scores for the two sets of subjects were in agreement using Spearmans rank correlation (ρ), a nonparametric measure for testing monotonic relationship between two variables. We computed the average correlation over 25 of these random split iterations, yielding an average correlation of $\rho = 0.76$. This high consistency in object memorability indicates that, like full images, object memorability is a shared property across subjects. People tend to remember (and forget) the same objects in images, and exhibit similar performance in doing so. Thus memorability of objects in images can potentially be predicted with high accuracy. In the next section, we study the various factors that drive object memorability in images.

3. Understanding Object Memorability

In this section, we aim to better understand how object memorability is influenced by visual factors that manifest themselves in natural images. Specifically, we study the relationship between simple color features, visual saliency, object semantics, and how memorable or forgettable an object in an image is to humans. The results of this study can be used to guide the development and innovation of automated algorithms that can predict object memorability.

3.1. Can simple features explain memorability?

While simple low-level image features are traditionally poor predictors of image memorability [18] (with good reason [27]), the question arises whether such features play any role in determining object memorability in images. To address this question and following a similar strategy as in [18], we compute the mean and variance of each HSV color channel for each object in our dataset, and compute the Spearman rank correlation with the corresponding object memorability score (refer to Figure 3). We see that the mean ($\rho = 0.1$) and variance ($\rho = 0.25$) of the V channel correlates weakly with object memorability, suggesting that brighter and higher contrast objects may be more memorable. On the other hand, essentially no relationship exists between memorability and either the H or S channels.

This deviates slightly from the findings in [18], which show mean hue to be weakly predictive of image memorability. This difference could be due to the fact that the dataset in [18] contains blue and green outdoor landscapes that are less memorable than the warmly colored human faces and indoor scenes. In contrast, outdoor scene-related segments such as sky and ground were not included as objects in our dataset. From these results, we see that, like image memorability, simple pixel statistics do not play a significant role in determining object memorability in images.

3.2. What is the role of saliency in memorability?

Intuitively, we expect that objects within an image that are most salient are likely to be remembered, since they tend to draw a viewer's attention, i.e. a majority of his/her eye fixations will lie within those object regions. On the other hand, it is conceivable that some visually appealing regions will not be memorable, especially since aesthetic images are known to be less memorable [18]. When can visual saliency predict object memorability and what are the possible differences between the two? Studying the relationship between saliency and memorability is paramount for understanding object memorability in greater depth.

To address this query, we utilize the eye fixation data made available for the PASCAL-S dataset [30]. First, we compute the number of unique fixation points within the image segment of each object and the correlation between this metric and the object's memorability score (refer to Figure 4 (*left*)). We find this correlation to be positive and considerably high ($\rho = 0.71$), suggesting that fixation count and visual saliency may drive object memorability considerably. However, the large concentration of points on the bottom left part of the scatter plot in Figure 4 (*left*) suggests that part of the reason for this high correlation is that objects that have not been viewed (i.e. no fixation points associated with them) at all have essentially no memorability, and

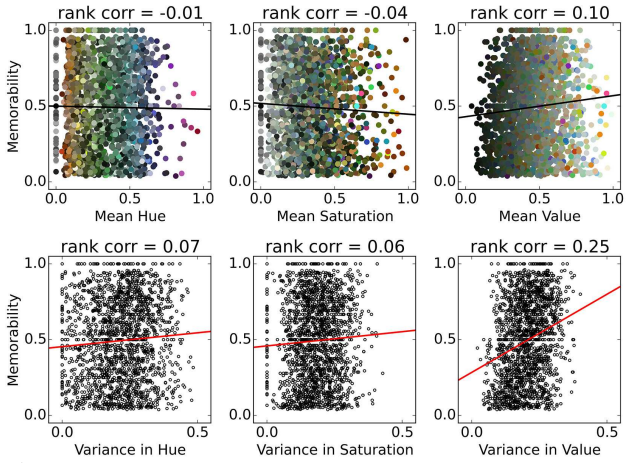

Figure 3: **Simple color features do not explain object memorability.** Correlations of object memorability scores with hue and saturation are near zero. Only value shows a weak correlation.

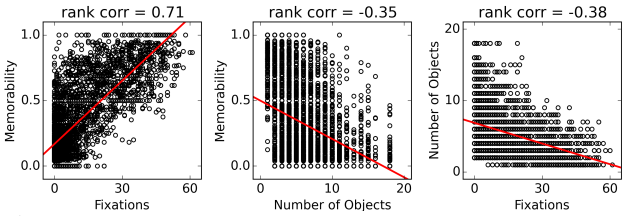

Figure 4: **Correlations between memorability, fixation count, and number of objects.** Left: Memorability and fixation counts correlate positively. Middle: Memorability and number of objects are negatively correlated. Right: Fixations and object counts are weakly negatively correlated.

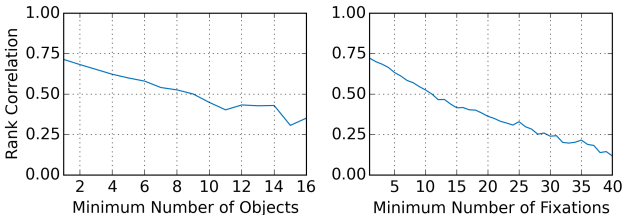

Figure 5: Correlation between object memorability and object fixation count as a function of minimum number of objects (left) and minimum number of fixations (right).

therefore will always imply correlation. If we remove these simple cases, we can examine whether or not the full range of memorability scores is predicted by fixation count. To investigate this, we plot the change in correlation between object memorability and fixations as the minimum number of fixations inside objects increases. For each minimum fixation count, we compute the memorability-fixation correlation again but *only* using objects that contain at least this number of fixations (refer to Figure 5 (*right*)). The decreasing trend in correlation indicates that as the number of fixations inside an object increases, the predictive ability diminishes significantly, indicating that the full range of memorability scores are not well predicted. In addition, Figure 5 (*left*) plots this correlation as a function of total number of objects in an image. Interestingly, as the number of objects in an image increases, the correlation between saliency, i.e. number of fixations, and memorability decreases sharply. The two remaining scatter plots in Figure 4 (*middle*) and (*right*) provide additional clues about the relationship between memorability and fixation count. Note that object count is negatively correlated with both memorability and fixation count. This makes sense, since people have more to look at in an image when more objects are present. In this case, they tend to look less at any single object, especially if some of these objects compete for saliency, and therefore may have a more difficult time remembering those objects.

In summary, saliency is a surprisingly good predictor of object memorability in simple contexts where few objects exist in an image or when an object has few interesting points, but it is a much weaker predictor of object memorability in complex scenes containing multiple objects that have many points of interest (refer to Figure 6).

Center Bias: Figure 7 illustrates another example where saliency and memorability diverge. Previous studies related

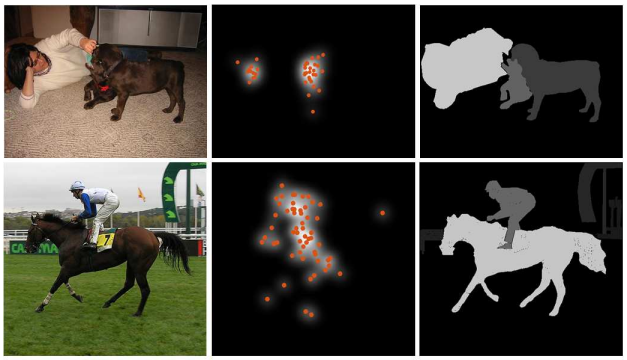

Figure 6: **Memorability prediction by saliency in complex scenes.** Top row: the memorability of the dog is low even though many humans fixate on it. Bottom row: Humans look at the person more than the horse although the horse is more memorable than the person.

to visual saliency have shown that saliency is heavily influenced by center bias [21, 40], primarily due to photographer bias (also evident in Figure 7 (*left*)) and viewing strategy [38]. Since our data collection experiment tries to control for the viewing strategy, memorability exhibits comparatively less center bias than saliency. This is most apparent when considering the difference in the solid ellipse in the right plot (shows where 95% of fixations are located), and the dashed ellipse (shows where 95% of the above-median memorable objects are located).

To the best of our knowledge, this work is the first to give an in-depth study of the relationship between saliency and memorability and to highlight how the two phenomena differ from each other.

3.3. How do object categories affect memorability?

In the previous section, we explored the relationship between visual saliency and object memorability. Now, we explore how an object's category influences the probability that it will be remembered.

3.3.1 Are some object categories more memorable?

For this analysis, we had three in-house annotators manually label the object segmentations in our dataset. The annotators were provided the original image (for reference) and the object segmentation and asked to assign a single category to the segment out of 7 possible categories: animal, building, device, furniture, nature, person, and vehicle. We chose these categories so that a wide range of object classes could be covered. For example, category "device" includes objects like utensils, bottles, and televisions, while "nature" includes objects like trees, mountains, and flowers etc. Figure 8 shows the distribution of the memorability scores of all 7 object categories in our dataset.

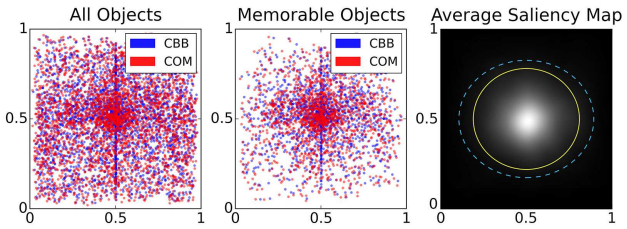

Figure 7: **Memorable objects and fixation locations.** Left: Normalized object locations for entire image data set. Both center of object bounding boxes (CBB, blue) and object center of mass (COM, red) are shown. Middle: Locations for memorable objects only. Right: Average ground truth saliency map across the entire dataset. The solid yellow line marks the region with 95% of all normalized fixation locations. The dashed blue line marks the region with above-median memorable objects. Center bias is more strongly expressed in the fixation locations.

Results in Figure 8 give a sense of how memorability changes across different object categories. Animal, person, and vehicle are all highly memorable classes, each associated with an average memorability score greater than or close to 0.5. Interestingly, all other categories have an average memorability lower than 0.25, indicating that humans do not remember objects from these categories very well. In particular, furniture is the least memorable category with an average score of only 0.14. This is possibly due to the fact that most objects in the furniture, nature, and building categories either appear mostly in the background or are occluded, which likely decreases their memorability significantly. In contrast, objects from the animal, person, and vehicle categories appear mostly in the foreground, leading to a higher memorability score on average. Interestingly, the most memorable objects from building, furniture, and nature tend to have an average memorability in the range of $0.4 - 0.8$, whereas the score of the most memorable objects from person, animal and vehicle is higher than 0.9. While the differences in the memorability of different object categories could be driven due to factors like occlusion, size, background/foreground, or photographic bias, the distribution in Figure 8 suggests that humans remember some object classes such as person, animal, and vehicle irrespective of external nuisance factors and these categories are *intrinsically* more memorable than others.

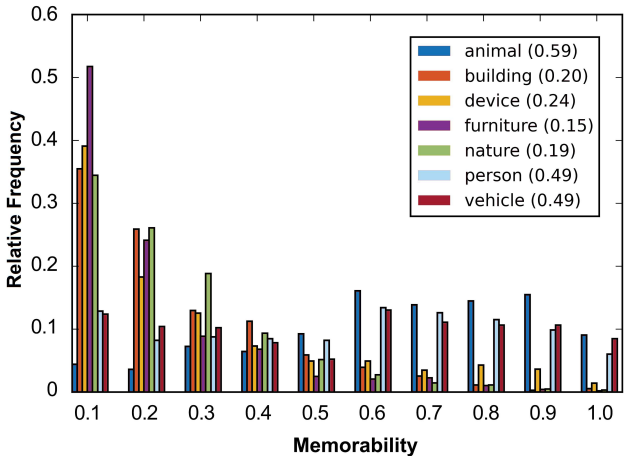

Figure 8: **Some object categories are more memorable than others.** Categories like furniture, nature, building, and device tend to have a large majority of objects with very low memorability scores. Objects belonging to animal, person, and vehicle categories are remembered more often.

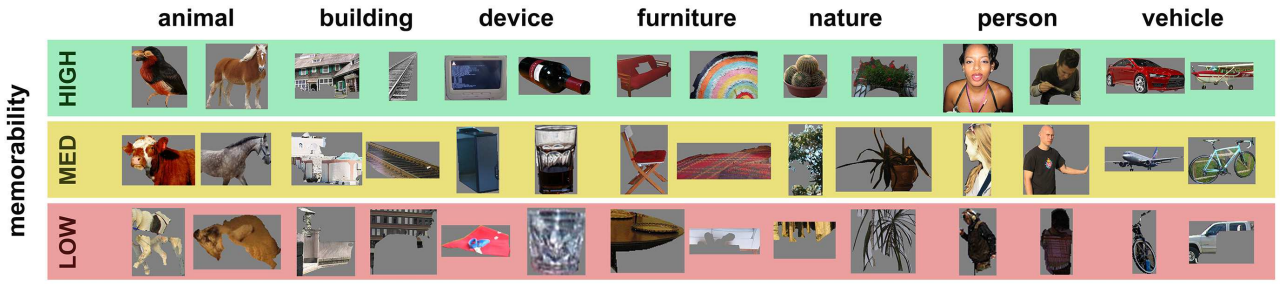

Figure 9: **Memorability of object categories.** Most memorable, medium memorable and least memorable objects from each of the 7 categories.

3.3.2 Exploring category-specific memorability

As demonstrated above, some object categories (i.e. animal, person, and vehicle) tend to be more memorable than others. However, not all objects in the same category are equally memorable. The examples in Figure 9 show the most memorable, medium memorable, and least memorable objects for each category. Across categories, medium to high memorable objects tend to have little to no occlusion. However, less memorable objects tend to be those that are occluded and obstructed by other objects. What other category-related factors could influence the memorability of objects? Among the possible factors, we explore how category-specific object memorability is influenced by **(i)** the number of objects in an image and **(ii)** the presence of other object categories.

Number of objects: Figure 10 shows the change in average memorability for the different categories when the minimum number of objects within an image is increased. Results indicate that the number of objects present in an image is an important factor in determining memorability. For example, as the number of objects in an image increases, the memorability of animals and vehicles decreases sharply, most likely as a result of competition for attention. Although the memorability of vehicles starts to show a slight increase for objects greater than 8, this arises only due to insufficient data (number of images is less than 30). Interestingly, the memorability of the person category does not change significantly when an increasing number of objects exist in the image. This suggests that people are not only one of the most memorable object categories, but that their memorability is the least sensitive to the presence of object clutter in an image.

Inter-category memorability: How much is the memorability of a particular object category affected when it co-occurs with another object category (or another instance of the same category)? To quantify the effect of one category on another, we consider each pairwise combination of categories and gather all images that contain at least one object from both categories. By taking one category as the *reference* and the other as the *distractor*, we compute the average memorability score $m_{R|D}$ of the reference in the images common to the reference and distractor. To isolate the

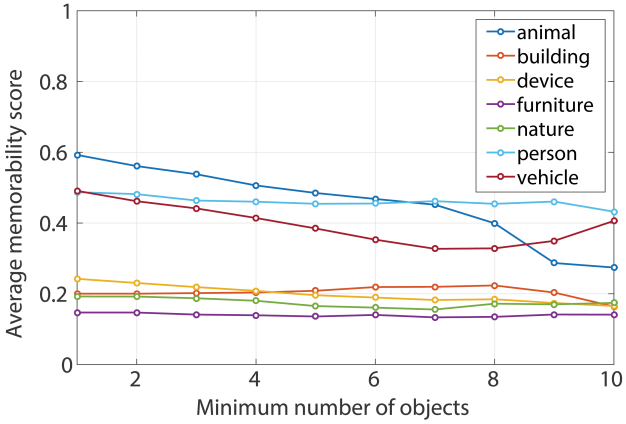

Figure 10: **Object number affects category-specific memorability.** For each category, a curve is plotted that shows the change in average memorability with an increase in the number of objects. The memorability of objects belonging to categories like animals and vehicles goes down significantly with an increase in object number.

effect of the distractor, we compute the memorability difference $\Delta m = (m_{R|D} - m_R)$, where m_R is the memorability score of the reference in all images where it exists. Figure 11 shows Δm for all possible reference and distractor pairs. It is clear that Δm for low-memorability categories (i.e. nature, furniture, device, and building) is not significantly affected by the presence of other categories.

Also, the memorability of the animal category maintains its high score in the presence of other categories, except vehicles, people, and itself, where it decreases substantially.

Figure 11: **Inter-category object memorability relationship.** Effect of distractor categories on the memorability of reference categories

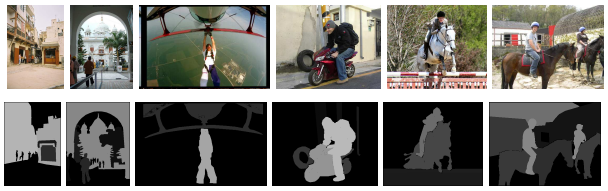

Figure 12: **Memorability of people in presence of other categories.** Top row: Images where a person co-occurs with other categories. Bottom row: Ground truth object memorability maps. In the presence of buildings, the memorability of person can drop. In the presence of a vehicle or animal, the person is usually more memorable.

The memorability of people tends to be unaffected by the presence of most other categories including itself. However, it decreases in the presence of vehicles and buildings. This could be due to the fact that people in images containing vehicles or buildings are usually zoomed out and smaller in size (refer to Figure 12). The memorability of the vehicle category is strongly affected by the presence of other object categories. In particular, it drops significantly in the presence of other vehicles, people, and animals.

In summary, when an animal, vehicle, or person co-occur in the same image, the memorability of all three categories usually decreases. However, this pattern of change in memorability is category-specific in general. For example, when a vehicle and animal are present in the same image, the animal is generally more memorable, even though both their memorability scores drop significantly. When a vehicle or an animal co-occurs with a person, the person is generally more memorable (also shown in Figure 12).

3.4. How are object & image memorability related?

Until now, we have studied what objects people remember and the factors that influence their memorability, but to what extent does the memorability of individual objects affect the overall memorability of an image? Moreover, if an image is highly memorable, what can we say about the memorability of the objects inside those images (and vice versa)? To shed light on these queries, we conducted a second large-scale experiment on Amazon Mechanical Turk for all images in our dataset to gather their respective *image* memorability scores. For this experiment, we followed the same strategy as the memory game experiment proposed in [23]. A series of images from our dataset and Microsoft COCO dataset [31] (i.e. 'filler' images) were flashed for 1 second each, and subjects were instructed to press a key whenever they detected a repeat presentation of an image. A total of 350 workers participated in this experiment with each image being viewed 80 times on average. The rank correlation after averaging over 25 random splits was 0.7, determining consistency in the image memorability scores.

Using results from the previous experiments, we computed the correlation between the scores of the single most memorable *object* in each image and the memorability score of each *image*. This correlation is moderately high with

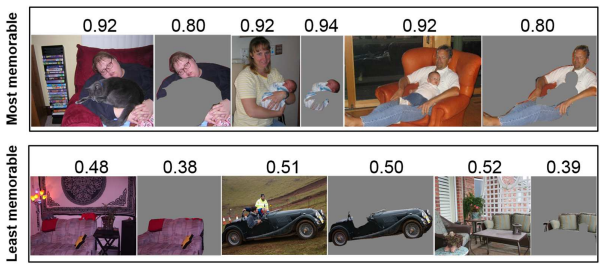

Figure 13: **Max object memorability predicts image memorability.** Top row: most memorable images taken from our dataset along with their highest memorable object and their respective memorability scores. Bottom row: least memorable images in the dataset along with their most memorable object and their respective memorability scores. We notice that maximum object memorability correlates strongly with image memorability in both cases.

$\rho = 0.4$, suggesting that the most memorable object in an image plays a crucial role in determining the overall memorability of an image. To investigate this finding in relation to some extreme cases, we repeated the same analysis as above but on a subset of the data containing the 100 most memorable images and the 100 least memorable images. The correlation between maximum object memorability and image memorability for this subset of images increased significantly to $\rho = 0.62$. This means that maximum object memorability serves as a strong indicator of whether an image is *highly* memorable or *not* memorable at all. In other words, images that are highly memorable contain at least one highly memorable object and images with low memorability usually do not contain a single highly memorable object (refer to Figure 13).

To study the effect of maximum object memorability across categories, we computed the correlation between maximum object and image memorability for each individual object category. The correlation for the categories were: animal ($\rho = 0.38$), building ($\rho = 0.22$), device ($\rho = 0.47$), furniture ($\rho = 0.53$), nature ($\rho = 0.64$), person ($\rho = 0.54$), and vehicle ($\rho = 0.30$) which shows that certain categories are more strongly correlated than others. For example, images containing animals, buildings, or vehicles as the most memorable objects tend to have varying degree of image memorability (indicated by their lower ρ values). On the other hand, device, furniture, nature, and person are strongly correlated with image memorability, indicating that if an image's most memorable object belongs to one of these categories, the object memorability score is strongly predictive of the image memorability score. We can imagine scenarios in which this information is potentially useful. For example, in vision systems that are tasked to predict scene memorability, a *single* object and its category can serve as a strong prior in predicting this score.

4. Predicting Object Memorability

This work makes available the very first dataset containing ground truth memorability of constituent objects from a

highly diverse image set. In this section, we show that our dataset can be used to benchmark computational memorability models and serve as a stepping stone in the direction of automated object memorability prediction.

Baseline models: As a first step, we propose a simple baseline model that utilizes a conv-net [28, 20] trained on the ImageNet dataset [37]. Since object categories play an important role in determining object memorability (Section 3.3), and deep learning models have recently been shown to achieve state-of-the-art results in various recognition tasks, including object recognition [14, 29], we believe that this simple model can serve as an adequate baseline for object memorability prediction. We first generated object segments by using MCG, a generic object proposal method proposed in [2]. Next, we trained a support vector regressor (SVR) using 6-fold cross-validation on the original object segments to map deep features to memorability scores. We used this model to predict memorability scores for the top $K = 20$ object segments obtained using the MCG algorithm. After predicting these memorability scores, memorability maps were generated by averaging the scores of these top K segments at the pixel level. Since image features like SIFT [33] and HOG [10] have previously been shown to achieve good performance in predicting image memorability [18], we built a second baseline model using these features for comparison. Training and testing of this model was performed similar to the conv-net model.

Evaluation: To evaluate the accuracy of the predicted object memorability maps, we computed the rank correlation between the mean predicted memorability score inside each of the object segments and their ground truth memorability scores. These results are reported in Figure 14. Clearly, the conv-net baseline, DL-MCG, performs considerably well ($\rho = 0.39$). In contrast, the baseline trained using HOG and SIFT, H+S, achieves a much lower performance ($\rho = 0.27$). Saliency maps generated from saliency algorithms are also likely to have some degree of overlap with memorability and are therefore worth comparing to our baseline, especially given the absence of other memorability prediction methods. To this end, we included 8 state-of-the-art saliency methods (top performing methods according to benchmarks in [5, 4]): GB [15], AIM [7], DV [16], IT [19], GC [9], PC [35], SF [36], and FT [1] to our comparison. Figure 14 shows that the H+S baseline is outperformed by most saliency methods. Thus, even though models using SIFT and HOG have previously demonstrated high predictive power for image memorability, they may not be as well suited for the task of predicting object memorability. The deep-net baseline model, DL-MCG performs better than all other saliency methods with only PC ($\rho = 0.38$), SF ($\rho = 0.37$), and GB ($\rho = 0.36$) showing comparable performance. A common factor between these saliency methods is that they explicitly add center bias to their implementation. Although object memorability exhibits less center bias when compared to eye fixations, it still tends to be biased somewhat towards the center due to photographer bias (see Section 3.2), which could be a reason for the high performance of these saliency methods. While DL-MCG performed favorably in predicting object memorability, its accuracy is highly dependent on the quality of the segmentations used. To illustrate this fact, we redo the prediction task but with the ground truth segments replacing the MCG segments. The resulting baseline is referred to as DL-UL, which can be considered the gold standard or the upper bound on automated object memorability prediction. Its correlation score is very high and close to human performance ($\rho = 0.7$), which suggests that the conv-net model does have high predictive ability but that it is sensitive to the image segmentations it is applied to.

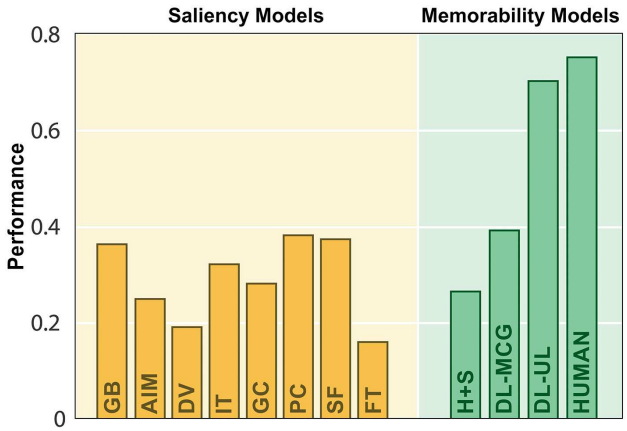

Figure 14: **Rank correlation of predicted object memorability.** Accuracy of the baseline and saliency algorithms on proposed benchmark.

5. Conclusion

In this paper, we propose the problem of understanding the memorability of objects in images. To this end, we obtained ground truth data that helps to study and analyze this problem in depth. We show that the category of an object has meaningful influence on its memorability, and that visual saliency can predict object memorability to some degree. Moreover, we studied the relationship between image and object memorability and compiled a benchmark dataset for automated object memorability prediction. Future work will involve studying the influence of non-object image regions and scene context on memorability.

Acknowledgement. This work was supported by competitive research funding from King Abdullah University of Science and Technology (KAUST). M.-H. Yang is supported in part by NSF CAREER Grant (No.1149783) and NSF IIS Grant (No.1152576). Special thanks to James C. McEntire for helping with graphics.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages

1597–1604, 2009. 8

[2] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, pages 328–335, 2014. 8

[3] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569, 2012. 1

[4] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, pages 414–429. 2012. 1, 8

[5] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *TIO*, 22(1):55–69, 2013. 8

[6] 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. 3

[7] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, 2005. 1, 8

[8] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 2015. 1, 2

[9] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011. 8

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 8

[11] A. Dave, R. Dubey, and B. Ghanem. Do humans fixate on interest points? In *ICPR*, pages 2784–2787, 2012. 1

[12] R. Dubey, A. Dave, and B. Ghanem. Improving saliency models by predicting human fixation patches. In *ACCV*, pages 330–345, 2014. 1

[13] M. Everingham and J. Winn. The pascal visual object classes challenge 2010 (voc2010) development kit, 2010. 2

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 8

[15] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 1, 8

[16] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, pages 681–688, 2009. 1, 8

[17] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, pages 2429–2437, 2011. 1

[18] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *PAMI*, 36(7):1469–1482, 2014. 1, 2, 3, 4, 8

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998. 1, 8

[20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*, pages 675–678, 2014. 8

[21] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 5

[22] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva. Modifying the memorability of face photographs. In *ICCV*, pages 3200–3207, 2013. 1

[23] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *ICCV*, 2015. 7

[24] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva. Image memorability and visual inception. In *SIGGRAPH Asia Technical Briefs*, 2012. 1

[25] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *NIPS*, pages 305–313, 2012. 1, 2

[26] J. Kim, S. Yoon, and V. Pavlovic. Relative spatial features for image memorability. In *MM*, pages 761–764, 2013. 2

[27] T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3):558, 2010. 3

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 8

[29] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, pages 609–616, 2009. 8

[30] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 2, 4

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. 2014. 3, 7

[32] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *PAMI*, 33(2):353–367, 2011. 3

[33] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 8

[34] M. Mancas and O. Le Meur. Memorability of natural scenes: the role of attention. In *ICIP*, 2013. 2

[35] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, pages 1139–1146, 2013. 8

[36] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 8

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision*, 2015. 8

[38] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of vision*, 9(7):4, 2009. 5

[39] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 3

[40] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32, 2008. 5