Image memorability and visual inception

Citation: Khosla, Aditya, Jianxiong Xiao, Phillip Isola, Antonio Torralba, and Aude Oliva. "Image Memorability and Visual Inception." SIGGRAPH Asia 2012 Technical Briefs on - SA '12 (2012), SIGGRAPH Asia 2012, November 28-December 1, 2012, Singapore. Article no. 35: p.1-4.

As Published: http://dx.doi.org/10.1145/2407746.2407781

Publisher: Association for Computing Machinery

Persistent URL: http://hdl.handle.net/1721.1/90955

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Image Memorability and Visual Inception

Aditya Khosla* Jianxiong Xiao* Phillip Isola Antonio Torralba Aude Oliva
Massachusetts Institute of Technology

Figure 1: *Sample of the database used for the memory study. The images are sorted from more memorable (left) to less memorable (right).*

Abstract

When glancing at a magazine, or browsing the Internet, we are continuously being exposed to photographs. However, not all images are equal in memory; some stitch to our minds, while others are forgotten. In this paper we discuss the notion of image memorability and the elements that make it memorable. Our recent works have shown that image memorability is a stable and intrinsic property of images that is shared across different viewers. Given that this is the case, we discuss the possibility of modifying the memorability of images by identifying the memorability of image regions. Further, we introduce and provide evidence for the phenomenon of *visual inception*: can we make people believe they have seen an image they have not?

Keywords: memorability, computational photography, image processing, image property, smart image editing

1 Introduction

When glancing at a magazine or browsing the Internet we are continuously exposed to photographs and images. Despite this overow of visual information, humans are extremely good at remembering thousands of pictures and a surprising amount of their visual details [Brady et al. 2008; Konkle et al. 2010; Standing 1973]. While some images stick in our minds, others are ignored or quickly forgotten. Artists, advertisers, and photographers are routinely challenged by the question 'what makes an image memorable?' and are then presented with the task of creating an image that will be remembered by the viewer.

While psychologists have studied human capacity to remember visual stimuli [Brady et al. 2008; Konkle et al. 2010; Standing 1973], little work has systematically studied the differences in stimuli that make them more or less memorable. In a recent paper [Isola et al. 2011b], we quantified the memorability of 2222 photographs as the rate at which subjects detect a repeat presentation of the image a few minutes after its initial presentation. The memorability of these images was found to be consistent across subjects and across a variety of contexts, making some of these images intrinsically more memorable than others, independent of the subjects past experiences or biases. Thus, while image memorability may seem like a quality that is hard to quantify, our recent work suggests that it is not an inexplicable phenomenon.

Being an intrinsic property, the memorability of an image can potentially be modified by changing certain properties or elements contained in an image. In order to do this, we first need to understand the characteristics of an image that affect its memorability. We [Isola et al. 2011a] show that image memorability cannot be predicted by typical attributes used by humans to describe images such as 'unusualness', 'aesthetically pleasing' or 'funny', and further that humans are no better than random number generators at predicting the memorability of an image. However, computer vision algorithms have been shown to be fairly effective at this task.

Our recent work [Khosla et al. 2012] extends this notion further to automatically determine the contribution of specific image regions to the memorability of an image. By modifying regions of high or low memorability we can potentially change the overall memorability of an image. While it may be possible to increase the memorability of an image by introducing memorable elements out of context, it is much more interesting and difficult to synthesize an image that is both realistic and more memorable than the original image. This could have far-reaching applications in various domains ranging from advertising and gaming (e.g. making logos more memorable), to education (e.g. if we could modify an image without losing it's main point) and social networking.

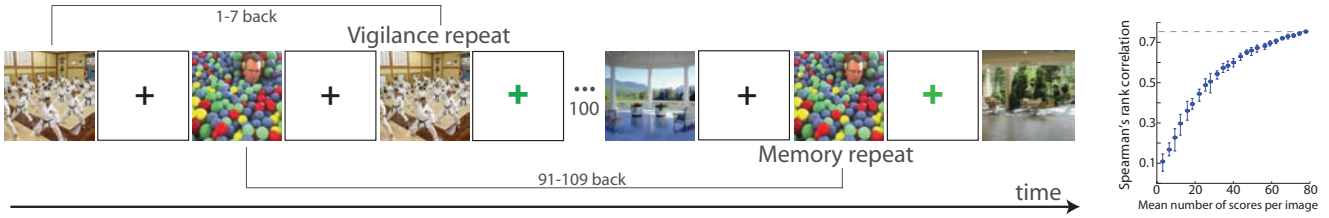
We take a step further and introduce the idea of *visual inception*: can we make people believe they have seen an image before that they have not? How much can we change an image while still making people believe they have seen it (i.e. the original image)? We provide evidence that suggests that this phenomenon is already observed, and it may be possible to do visual inception.

In this paper, we provide an overview of the recent works on image memorability [Khosla et al. 2012; Isola et al. 2011a; Isola et al. 2011b] that show that memorability is a measurable and intrinsic property of images (Sec. 2). Then we attempt to better understand the elements of an image that affect its memorability (Sec. 3). Further, we introduce the ideas of modifying memorability of images (Sec. 4) and visual inception (Sec. 5) and conclude in Sec. 6.

2 Measuring Memorability

Are images remembered by one person also more likely to be remembered by someone else? In this section, we characterize the consistency of image memory. In order to do so, we built a database (Fig. 1) depicting a variety of images, and we measured the probability of remembering each image over a large population of users.

*-indicates equal contribution



(a) The memory game (b) Human consistency

Figure 2: ***(a)*** *Mechanical Turk workers played a 'Memory Game' in which they watched for repeats in a long stream of images (Sec. 2.1). Vigilance repeats were used to ensure the workers were paying attention, resulting in a high quality of annotation.* ***(b)*** *Consistency across independent groups of people, measured by Spearman's rank correlation, increases as the number of scores per image is increased (Sec. 2.2).*

2.1 Memory Game

In order to measure image memorability, we presented workers on Amazon Mechanical Turk with a Visual Memory Game [Isola et al. 2011b]. In the game, participants viewed a sequence of images, each of which was displayed for 1 second, with a 1.4 second gap in between image presentations (Fig. 2(a)). Their task was to press the space bar whenever they saw an identical repeat of an image at any time in the sequence. Image sequences were broken up into levels of 120 images each. A total of 665 workers performed the task, completing an average of 13 levels per worker.

Unbeknownst to the participants, the sequence of images was composed of 'targets' (2222 images) and 'fillers' (8220 images) obtained from the SUN dataset [Xiao et al. 2010]. All images were scaled and cropped about their centers to be 256x256 pixels. The role of the fillers was two-fold: first, they provided spacing between the first and second repetition of a target; second, responses on repeated fillers constituted a vigilance task that allowed us to continuously check that participants were attentive to the task. Stringent criteria were used to continuously screen worker performance.

After collecting the data, we assigned a 'memorability score' to each image, defined as *the percentage of correct detections by participants*. On average, each image was scored by 78 participants. The average memorability score was 67.5% with a false alarm rate of 10.7%. As the false alarm rate was low in comparison with correct detections, correct detections are unlikely to be lucky confusions. Therefore, we believe our memorability scores are a good measure of correct memories.

2.2 Consistency Analysis

To evaluate human consistency, we split our participant pool into two independent halves, and quantified how well image scores measured on the first half of the participants matched image scores measured on the second half of the participants (Fig. 2(b)). Averaging over 25 random split half trials, we calculated a Spearman's rank correlation (ρ) of 0.75 between these two sets of scores. We found that the rank correlation between different group of participants increased as we increased the number of scores per image. This level of consistency suggests that information intrinsic to the images is likely used by different people to remember them. Overall, our experiments provide strong evidence that image memorability is an intrinsic property of images.

3 Predicting Memorability

Among the many reasons why an image might be remembered by a viewer, we investigate the following factors: simple image features, object and scene semantics, computer vision features, semantic attributes and human estimation. As described in Sec. 2.1, we compute the average Spearman's rank correlation (ρ) across 25 random splits of the data to compare different methods. We trained a ranking support vector machine [Joachims 2006] to map different features to memorability scores, using half the images for training and the other half for testing in each trial. The results are summarized in Fig. 3.

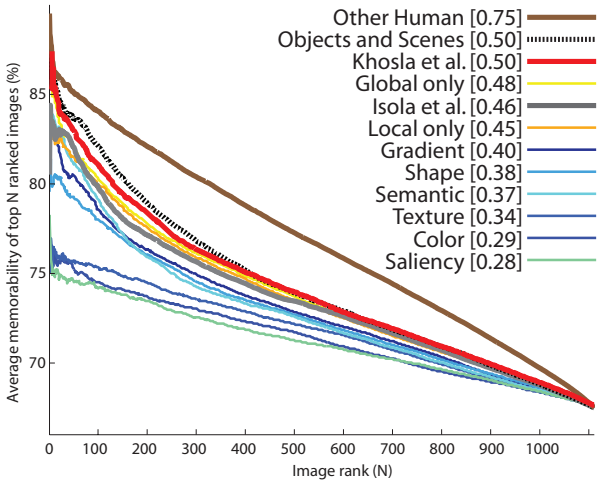


Figure 3: *Memorability prediction: Images are ranked by predicted memorability and plotted against the cumulative average of measured memorability scores. Saliency, color, texture, semantic, shape and gradient refer to typical computer vision features, and 'Isola et al.' is the performance when combining these features (Sec. 3). 'Khosla et al.' refers to the method using image regions, combining 'Local only' and 'Global only' features (Sec. 4.1), while 'Objects and Scenes' refers to using ground truth annotation.*

Simple image features: These consist of simple image statistics such as mean of individual color channels, mean hue/saturation and mean intensity. These features exhibited very weak rank correlation ($|\rho| < 0.1$) with memorability scores.

Object and scene semantics: Using LabelMe [Russell et al. 2008], each image in our target set was segmented into object regions and each of these segments was given an object class label by a human user (e.g person, mountain, stethoscope). This gave us object semantics such as 'object counts', 'object areas', and object presence. Combining this information with scene category label [Xiao et al. 2010; Ehinger et al. 2011] led to a relatively good set of features for predicting memorability ($\rho = 0.50$). While effective, it is expensive to obtain such detailed annotation for large-scale datasets.

Computer vision features: A variety of computer vision features (described in [Khosla et al. 2012]) corresponding to color, shape, texture and gradient are automatically extracted from each image. These features are found to be relatively effective at predicting memorability with 'gradient' giving a maximum performance of $\rho = 0.40$ for any individual feature. Combining these features leads to an improved performance of 0.46. This correlation is less

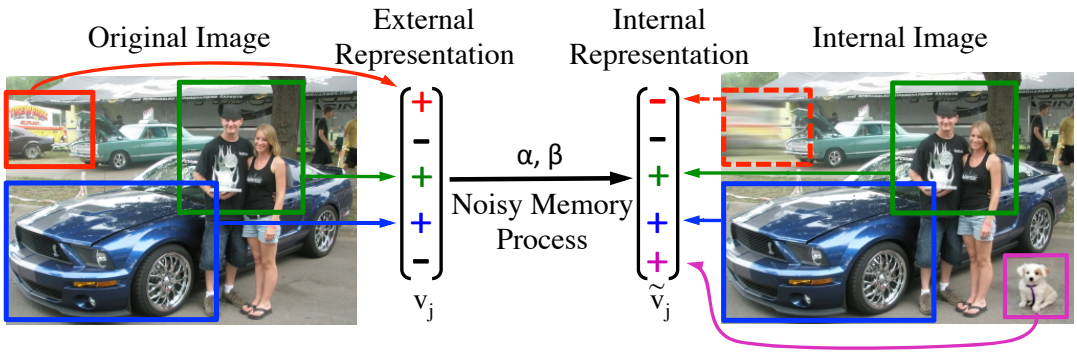


Figure 4: *Overview of our probabilistic framework. This figure illustrates a possible external or 'observed' representation of an image. The conversion to an internal representation in memory can be thought of as a noisy process where some elements of the image are changed probabilistically as described by α and β (Sec. 4.1). The image on the right illustrates a possible internal representation: the **green** and **blue** regions remain unchanged, while the **red** region is forgotten and the **pink** region is hallucinated. Note that the internal representation/image cannot be observed and is only shown here for illustrating the framework.*

than human predictions, but close to our best predictions from labeled annotations.

Semantic attributes and human estimation: We obtained various attribute annotations about spatial layout, aesthetics, emotions, dynamics and people such as 'Is aesthetic?', 'Exciting?' and 'Making eye contact?' [Isola et al. 2011a]. Contrary to popular belief, unusual and aesthetically pleasing images are not predominantly the most memorable ones. having a negative correlation ($\rho < -0.1$) with memorability. Overall, we found that images of enclosed spaces containing people with visible faces are memorable, while images of vistas and peaceful settings are not.

Further, we asked humans to estimate whether they were likely to remember an image the following day. We found that the human estimate was negatively correlated ($\rho = -0.19$) with the true memorability of images, suggesting that automatic algorithms are essential for this task.

4 Modifying Memorability

We propose the idea of modifying image memorability and describe our current work in this direction. Modifying memorability can be thought of a process consisting of two main steps, (1) identifying the memorability of different image regions, and (2) modifying the appropriate image regions based on requirements such as increasing or decreasing memorability, malleability (i.e. the proportion of an image that should remain unchanged, and whether new objects should be added or not) or an image mask (i.e. specify regions that should not be modified).

One possible approach to identify memorability of different image regions is to manually segment images into scenes and objects. However, this method suffers from several shortcomings. First, it is difficult to scale to large datasets, which have been shown to be important for modifying images while making them look realistic [Hays and Efros 2007]. Second, it is difficult to determine *a priori* the granularity of the segmentation e.g. should a person form just one image segment or should his arms, legs, torso and face be individual segments. We overcome these limitations by combining local and global features in an interpretable model [Khosla et al. 2012], to build an automatic memorability map and provide a strong model for predicting image memorability automatically.

4.1 Memorability of Image Regions

We propose to predict memorability using the process of forgetting different image regions, as illustrated in Fig. 4. The external representation refers to the original image which is shown to an observer, while internal representation refers to the noisy representation of the same image that is stored in the observers memory. We model the image as a combination of different types of image regions and features. After a delay between the first and second presentation of an image, people are likely to remember some image regions and objects more than others. For example, as shown in [Isola et al. 2011b], people and close up views on objects tend to be more memorable than natural objects and regions of landscapes, suggesting for instance that an image region containing a person is less likely to be forgotten than an image region containing a tree. We postulate that the farther the stored internal representation of the image is from its veridical representation, the less likely it is to be remembered.

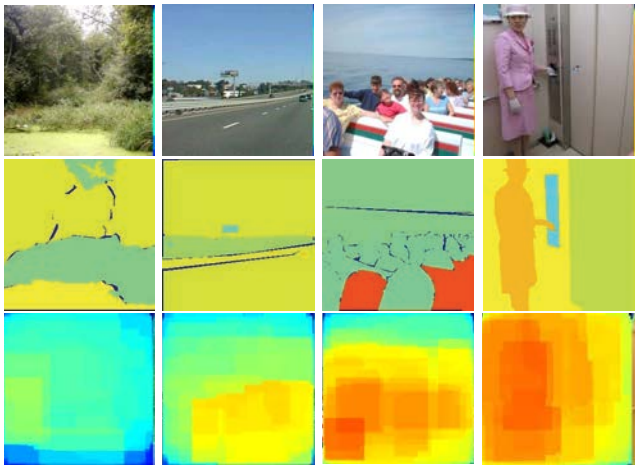


Figure 5: *Memorability maps of images (top row) using ground truth segments (middle row), and automatic methods (bottom row). The images are arranged from less memorable (left) to more memorable (right).*

We model this forgetting process in a probabilistic framework (Fig. 4). We assume that the representation of an image is composed of image regions where different regions of an image correspond to different sets of objects. These regions have different probabilities of being forgotten and some regions have a probability of being imagined or hallucinated. The likelihood of an image to be remembered depends on the distance between the initial image representation and its internal degraded version. An image with a larger distance to the internal representation is more likely to be forgotten, thereby the image should have a lower memorability score. In our algorithm, we model this probabilistic process and show its effectiveness at predicting image memorability (Fig. 3) and at producing interpretable memorability maps (Fig. 5).

Fig. 3 shows that when combining the local information learned by the above model, with global information from standard computer vision features, we are able to achieve comparable performance to when ground truth image annotation (e.g. object segments and scene category) is used. Despite using the same set of features for both local and global information, we are able to obtain performance gain suggesting that our algorithm is effective at capturing local information in the image that was overlooked by the global features. Further, from Fig. 5 we observe that the automatically learned memorability maps are similar to those obtained using ground truth objects and segments.

5 Visual Inception

The concept of 'inception', while largely fictional for now, may not be as far fetched as it seems. Consider the following experiment: A participant is told that a set of images will be shown to him one after another, and his task is to memorize the content of these images. An image is shown for one second followed by a blank screen for another second, followed by another image and so on (similar to the Memory Game, Fig. 2(a)) with a total of 10 images. Now the participant is shown one image at a time and asked to press a key if he believes he has seen the image before.

While not done formally, the authors of this work have independently conducted this experiment in a number of talks with diverse audiences, with highly consistent results. In a talk, the audience is asked to clap their hand to signal that they have seen a particular image. Unlike the Memory Game, in this case, we show the participants one set of images (top row, Fig. 6) and test them with another set of images (bottom row, Fig. 6) that are of a different environment (i.e. have different set of objects and are at a different place) but have a similar overall structure, measured via GIST [Oliva and Torralba 2006]. Additionally, we include some images that are not similar to the ones shown, to ensure the audience is paying attention. Remarkably, we find that most members of the audience are convinced that they have seen the images before when they actually have not! The only way to make them believe otherwise is to show both sets of images side-by-side as shown in Fig. 6.

While human memory is great, it is not perfect. This provides evidence of the fact that it may be possible to do visual inception, or essentially *make people believe that they have seen a particular image they have not*. Let us assume that we have some original image I, and a modified version of an image I'. An important question is, how much can we modify image I while still making people believe that they have seen it, even though they have only observed I'? Taking a step farther, is it possible to change the memorability of I' as compared to I, and indirectly affect the memorability of I without modifying its content (as people would believe they have already seen it, even though they have only observed I').

6 Conclusion

In this paper, we discussed the concept of image memorability and have shown that there is a large degree of consistency among different viewers, and that some images are more memorable than others even when there are no familiar elements (such as relatives or famous monuments). Also, we found that memorability is not a typical image property and, contrary to popular belief, cannot be characterized by common attributes such as unusualness and aesthetic beauty. Further, we proposed the idea of modifying image memorability, composed of identifying memorability of different image regions and modifying them. We proposed a probabilistic framework to predict memorability of image regions automatically and demonstrated its effectiveness in producing interpretable memorability maps. Lastly, we introduced and gave evidence for the idea of visual inception. Future development of such automatic algorithms

Figure 6: *Visual inception experiment. Top row shows the set of images shown to participants, while bottom row shows the images they are queried with (Sec. 5). People tend to believe they have seen the bottom row of images, after being shown the first row even though a significant proportion of the image has been modified including the objects, viewpoint and environment.*

of image memorability could have many exciting and far-reaching applications in computer science, graphics, media, designs, gaming and entertainment industries in general.

Acknowledgements
This work is funded by NSF grant (1016862) to A.O, Google research awards to A.O and A.T, ONR MURI N000141010933 and NSF Career Award (0747120) to A.T. J.X. is supported by Google U.S./Canada Ph.D. Fellowship in Computer Vision.

References

BRADY, T. F., KONKLE, T., ALVAREZ, G. A., AND OLIVA, A. 2008. Visual long-term memory has a massive storage capacity for object details. *PNAS*.

EHINGER, K., XIAO, J., TORRALBA, A., AND OLIVA, A. 2011. Estimating scene typicality from human ratings and image features. In *CogSci*.

HAYS, J., AND EFROS, A. 2007. Scene completion using millions of photographs. In *TOG*, vol. 26, ACM, 4.

ISOLA, P., PARIKH, D., TORRALBA, A., AND OLIVA, A. 2011. Understanding the intrinsic memorability of images. In *NIPS*.

ISOLA, P., XIAO, J., TORRALBA, A., AND OLIVA, A. 2011. What makes an image memorable? In *CVPR*, 145–152.

JOACHIMS, T. 2006. Training linear SVMs in linear time. In *ACM SIGKDD*, 217–226.

KHOSLA, A., XIAO, J., TORRALBA, A., AND OLIVA, A. 2012. Memorability of image regions. In *NIPS*.

KONKLE, T., BRADY, T., ALVAREZ, G., AND OLIVA, A. 2010. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology*, 139 (3), 558–578.

OLIVA, A., AND TORRALBA, A. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research 155*, 23–36.

RUSSELL, B., TORRALBA, A., MURPHY, K., AND FREEMAN, W. 2008. LabelMe: a database and web-based tool for image annotation. *IJCV 77*, 1, 157–173.

STANDING, L. 1973. Learning 10000 pictures. *The Quarterly journal of experimental psychology 25*, 2, 207–222.

XIAO, J., HAYS, J., EHINGER, K., OLIVA, A., AND TORRALBA, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, IEEE.