Memorability of natural scenes: the role of attention

Matei Mancas, Olivier Le Meur

MEMORABILITY OF NATURAL SCENES: THE ROLE OF ATTENTION

Matei Mancas[*]

University of Mons - UMONS, Belgium
NumediArt Institute, 31, Bd. Dolez, Mons
matei.mancas@umons.ac.be

Olivier Le Meur[†]

University of Rennes 1, France
IRISA, Campus de Beaulieu, Rennes
olemeur@irisa.fr

ABSTRACT

The image memorability consists in the faculty of an image to be recalled after a period of time. Recently, the memorability of an image database was measured and some factors responsible for this memorability were highlighted. In this paper, we investigate the role of visual attention in image memorability around two axis. The first one is experimental and uses results of eye-tracking performed on a set of images of different memorability scores. The second investigation axis is predictive and we show that attention-related features can advantageously replace low-level features in image memorability prediction. From our work it appears that the role of visual attention is important and should be more taken into account along with other low-level features.

***Index Terms*—** Image memorability, Visual attention, Eye tracking, Inter-observer congruency, Saliency

1. INTRODUCTION

The study of images memorability in computer science is a recent topic [1, 2]. From those first attempts it appears that it is possible to predict the degree of picture's memorability quite well. Learning algorithms have been used to infer from a set of low-level visual features the extent to which a picture is memorable. Although Isola et al. [1] expressed the intuition that memorability and visual attention might be linked, they did not study further this relationship. Khosla et al. [2] proposed a local descriptor based on Itti's model [3]. The performance of this descriptor alone is low.

In this paper we intend to show that attention-based cues and features might have high importance in memorability both from a experimental and predictive point of views. In the next sections we will focus on an eye-tracking experiment using images from Isola's database and the cues which can be extracted from gaze behaviour and which might be related to the memorability score of the images. In section 3, we evaluate the relevance of two attention-related features and show that by using the same classifier we obtain comparable and even better memorability results than [1]. Finally, we discuss and conclude about the role of attention in memorability.

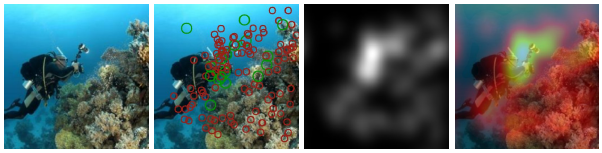


Fig. 1. (a) original pictures (memorability of 0.81 (high)); (b) fixation map (a green circle represents the first fixation of observers); (c) Saliency map and (d) heat map.

2. MEMORABILITY AND EYE-MOVEMENT

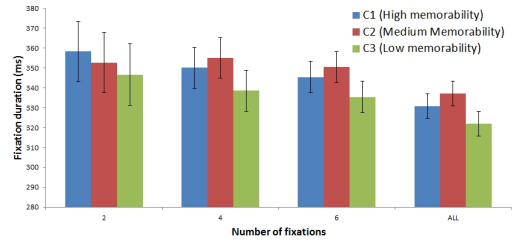
To shed light on the relationship between images memorability and visual attention, we conducted an eye-tracking experiment on images from the memorability database [1]. The eye-tracking data (images, fixations) used in this paper can be downloaded online at [4] or [5].

2.1. Method

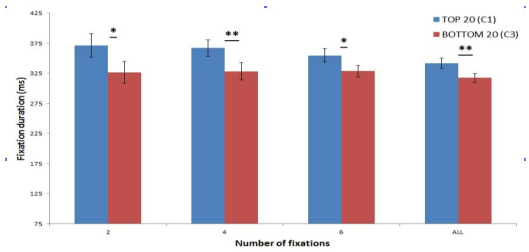
Participants and stimuli: Seventeen student volunteers (10 males, 7 females) with normal or corrected-to-normal vision took part to the eye tracking experiment. All were naïve to the purpose of the experiment and gave their full, informed consent to participate. We used 135 pictures extracted from [1] composed of 2222 pictures. Pictures are grouped in three classes of memorability (statistically significantly different), each composed of 45 pictures. The first class consists of the most memorable pictures ($C1$, score 0.82 ± 0.05), the second of typical memorability ($C2$, score 0.68 ± 0.04) and the third of the least memorable pictures ($C3$, score 0.51 ± 0.08).

Protocol: Pictures were displayed on a 19 inch monitor. The square images were centred on a white background, which filled the screen resolution of 800×600 pixels. At a viewing distance of 65 cm the stimuli subtended 17 degrees of visual angle. The eyes were tracked using the Face Lab 5 [6] with a sampling rate of 60Hz. Raw eye data were segmented into fixations and saccades by the Face Lab's system. The eye tracker

[*] Work performed while a research visit in IRISA Rennes.
[†] Both authors made equal contribution to this paper.



(a)



(b)

Fig. 2. (a) Fixation durations ($AVG \pm SEM$) for the 3 classes function of the first 2, 4, 6 and all the fixations; (b) Fixation durations for the most and less memorable pictures with the difference statistical significance (asterisks).

is calibrated using a 9-dot grid for each participant. Three sessions, each composed of 45 pictures randomly chosen were designed. Participants were instructed to look at the pictures given that they were required to answer a question at the end of each session to ensure that they were well involved in the exercise. Each picture was displayed for 5 seconds which is enough to catch the first impression involved in memorability. Pictures were separated by a blank image displayed for 2 seconds. The participants viewed the three classes in random order to avoid any bias in the final results.

2.2. Results

The analysis described below aims at proving that the visual behaviour of participants depends on the picture's memorability. We believe that attention is a step towards memory and therefore, this should influence the intrinsic parameters of eye movements such as the duration of visual fixations. Figure 1 illustrates this point. Two pictures are depicted. The first one has a memorability score of 0.81 whereas the second has a memorability score of 0.4. The average fixation durations for these two pictures are 391 and 278 ms, respectively. The average lengths of saccades are 2.39 and 2.99 degree of visual angle, respectively. In addition, if there is something in the picture that stand out from the background, the inter-observer congruency should be higher for memorable pictures. Results are presented in the following sections. This is the case for instance for the example presented on figure 1.

Fixation duration: Figure 2 illustrates the fixation durations (average (AVG) and standard error of mean (SEM)) for the three considered classes as a function of the viewing time. The fixation durations decrease with the degree of memorability of pictures, especially just after the stimuli onset. Fixations are the longest one when observers watch memorable pictures. A statistically significant difference is found between fixation durations when the top 20 most memorable and the bottom 20 less memorable are considered. This difference is confirmed for different viewing times.
These results are important since the duration of fixations reflects the deepness of the visual processing in the brain [7].
Inter-observer congruency: The congruency between observers watching the same stimulus indicates the degree of similarity between observers' fixations. A high congruency would mean that observers look at the same regions of the stimuli. Otherwise, the congruency is low. Generally the consistency between visual fixations of different participants is high just after the stimulus onset but progressively decreases over time [8]. To quantify inter-observer congruency, two metrics can be used: ROC [9] or a bounding box approach [10]. The former is a parametric approach contrary to the latter. The main drawback of the bounding box approach is its high sensitivity of outliers. A value of 1 indicates a perfect similarity between observers whereas the value 0 corresponds to the minimal congruency. Figure 3 shows the congruency as a function of viewing time (only the values obtained by the ROC-based metric are given but similar results are obtained by the second method). As expected the congruency decreases over time. Results also indicate that the congruency is highest on the class $C1$ (especially after the stimuli onset (first two fixations)). The difference between congruency of class $C1$ and $C2$ is not statistically significant. However, there is a significant difference between congruency of pictures belonging to $C1$ and $C3$. This indicates that pictures of classes $C1$ and $C2$ are composed of more salient areas which would attract more observer's attention.

These results show that memorability and attention are

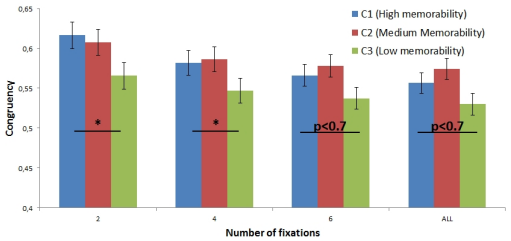
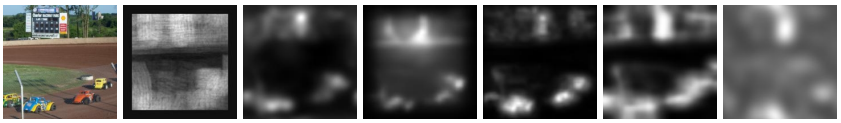


Fig. 3. Congruency as a function of viewing time. The symbol $*$ indicates that there is a significant statistically difference. Error bars correspond to the SEM.

linked. It would then be reasonable to use attention-based visual features to predict the memorability of pictures.



(a) Original (b) AIM (c) AWS (d) GBVS (e) RARE (f) Seo (g) LeMeur

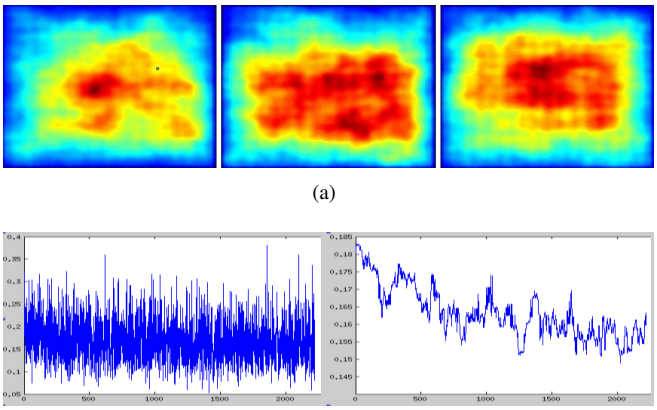
Fig. 4. (a) original pictures; (b) to (g) predicted saliency map from saliency models.

3. MEMORABILITY PREDICTION

Isola et al. showed that the best memorability prediction results are based on human labels containing information about the objects in the images. Nevertheless, these features are not available for any image and need time-consuming human annotations. Authors used then a mixture of several automatically extracted low-level features to approach the annotation-based results. The best result was achieved by mixing together GIST [11], SIFT [12], HOG [13], SSIM [14] and pixel histograms (PH). In this section we show that two other features of significantly smaller size which are related to attention can advantageously complement and replace some of the features proposed in [1]. For that purpose we use the SVR classifier and parameters from the code provided by [1].

Saliency map coverage: We extracted several times three classes of memorability composed of 45 images each randomly selected from a third of the most memorable images, a third of typical memorability and a third of low memorability images from the database proposed in [1]. Six state-of-the-art models of visual attention have been computed on those classes. Some saliency maps are displayed on Figure 4. From the saliency maps, the average saliency density is computed by accumulating the saliency maps of all the images within each class. The coverage (describes the spatial saliency density distribution) is here approximated by the mean of the normalized saliency maps. A low coverage would indicate that there is at least one salient region in the image. A high coverage may indicate that there is nothing in the scene visually important as most of the pixels are attended, but it might also indicate that there are several regions of interest which are randomly located on the images and the sum on the entire class covers most of the image. Figure 5 shows the saliency coverage of the RARE [15] model which is the one which is the most discriminant between the memorability classes. We computed this saliency coverage on several randomly generated classes (and show one of them on Figure 5) to be sure about the result reproducibility (this result is stable independently of the chosen images).

While the difference in terms of coverage between classes $C2$ and $C3$ is not obvious, this one is noticeable between the class $C1$ (the most memorable) and the two others. The class $C1$ coverage is lower which tends to show that there are mainly unique localized regions of interest while less memorable classes like $C2$ and $C3$ either do not have precise regions of interest or have several of those regions. The coverage of the RARE model saliency maps is thus used as a first feature in memorability prediction. Figure 5 (b) shows the result of the coverage for the whole database [1] from the less memorable to the more memorable image. The raw data (left plot of Figure 5(b)) is too noisy to be used alone (which is confirmed by the results in Table 1), but one can see on the median filtered version (right plot) that there is a negative correlation between the average coverage and memorability.



(a)

(b)

Fig. 5. Example of average coverage using RARE algorithm: (a) for classes $C1$, $C2$ and $C3$ (from left to right) on a random collection of 45 images out of the total of 2222 images; (b) from the less to the most memorable one on the whole database. Left: unfiltered data. Right: median filtered data.

Contrasted structures (visibility): A second feature used for memorability prediction is the contrast of the image structures. It is known [16] that object contrast is a strong attention feature. The most memorable images in Isola's database contain objects but also simpler backgrounds. This is especially true as the memorability score is established on the basis of a short observation time where complex backgrounds act like distractors and increase the visual masking phenomena.

To extract objects or at least structures contrast or "visibility" two approaches are used together (called V1 and V2). Both are based on low-pass filtering applied several times on images with kernels of increasing sizes like in Gaussian pyramids. The kernel sizes go from 3×3 kernels which eliminate some details to 80×80 which mainly result in very fuzzy images only providing a rough idea about their con-

Fig. 6. Low-pass filtering of images. From left to right: I_1, I_3, I_5, I_7 and I_9. RGB components are taken into account.

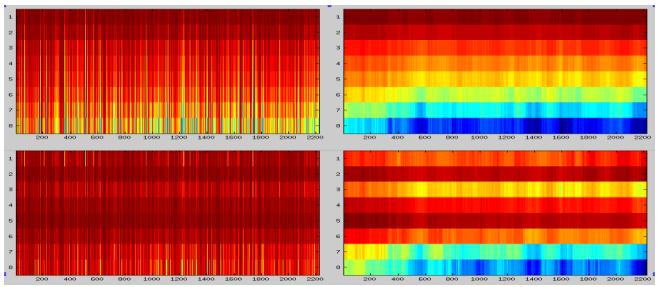


Fig. 7. Left: raw data for the 2222 images from the less memorable to the most memorable. Right: median-filtered data. First row: V1 data, second row: V2 data.

text or a gist. A set of 9 images I_i with $i \in \{1, 9\}$ where the first one ($i = 1$) is the original image and the last one ($i = 9$) is the most low-pass filtered. Figure 6 illustrates this approach on a given picture. To quantify the impact of low-pass filtering on the images, we measure their correlation ($corr$) after filtering: in the first approach (V1) the correlations between the initial image and all the others are computed: $\forall i \in \{2, 9\}, V1_i = |corr(I_1, I_i)|$. In the second approach (V2) the correlation between the successively filtered images are computed: $\forall j \in \{1, 8\}, V2_j = |corr(I_j, I_{j+1})|$. The correlation is the mean of the correlation of the RGB components. The main idea here is to see how an image reacts to multiple low pass filtering (which might be close to the forgetting process). Contrasted strong structures will be more resistant to low-pass filtering (higher correlation) while small details and structure with cluttered background will be much less resistant and achieve lower correlation scores. Figure 7 shows visibility feature vectors V1 and V2 computed for the whole 2222 images database. As in the previous section, the raw data both for V1 and V2 (left column of Figure 7) does not exhibit obvious differences. After median filtering (right column in Figure 7) differences between memorable and less memorable images are noticeable.

3.1. Results

The classifier and parameters are the ones from the code provided by [1]. Results shown in Table 1 are then perfectly comparable with the ones given in [1]. As already stated in sections 3, the proposed features are too noisy to provide good results if taken alone (see second and third column of Table 1). When combined to all of Isola et al. features but the GIST which is partially redundant with the visibility low-pass filtering of our V1 and V2 features, the result is comparable and even slightly better than the one of Isola et al. (Table 1). The proposed attention-related features are effective when taken together with other low-level features. It should also be noted that our features perform 2% better by using 17 dimensions instead of the 512 dimensions of the GIST feature which means 86% of the total features used by Isola et al. Table 2 shows the results where additional features from [1] were discarded. One by one, GIST and Pixels histograms, GIST and SIFT, GIST and HOG and GIST and SSIM were discarded from the features set. Still the results remain higher than the best combination of features in Isola et al. which shows the effectiveness of the proposed attention-related features even by replacing 1512 feature dimensions by 17 which means 58% only from the number of features in [1].

	Cov.	**Vis.**	**Best (No GIST)**	**Best Isola**
ρ	0.100	0.274	0.479	0.462

Table 1. Correlation results between the predicted memorability and labelled memorability. Column 2 and 3: proposed features alone (coverage, visibility). Column 4: proposed features and the SIFT, HOG, SSIM and Pixel histograms from [1], Column 5: Best feature-based combination from [1] (GIST, SIFT, HOG, SSIM, Pixel histograms).

	No Pixels	**No SIFT**	**No HOG**	**No SSIM**
ρ	0.476	0.474	0.470	0.468

Table 2. Correlation results obtained using the proposed features and combination of features excluding some features of [1] (no GIST and no Pixels histogram, no GIST and no SIFT, no GIST and no HOG, no GIST and no SSIM).

4. CONCLUSION

Isola et al. introduced an interesting approach to memorability but no relationship between attention and memorability was done. This paper shows that attention might play an important role in memorability both from an experimental and predictive perspectives when taken together with other features. The eye-tracking experiments made on a subset of the images database proposed by Isola et al. show that fixation duration and inter-observer congruency are interesting parameters well correlated with the images memorability. The prediction experiments made on the whole Isola et al. image database by using the same classifier, method and parameters showed that two attention-related features (RARE saliency map coverage and structures visibility) can advantageously replace some of the low-level features proposed in [1] and reduce in the same time the dimensionality of the feature set.

5. REFERENCES

[1] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 145–152.

[2] A. Khosla, J. Xiao, A. Torralba, and A. Oliva, "Memorability of image regions," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, December 2012.

[3] L. Itti, C. Koch, and E. Niebur, "A model for saliency-based visual attention for rapid scene analysis," *IEEE Trans. on PAMI*, vol. 20, pp. 1254–1259, 1998.

[4] M. Mancas, "Memorability project web page," `http://www.tcts.fpms.ac.be/attention/?categorie26/images-memorability`.

[5] O. Le Meur, "Home page," `http://people.irisa.fr/Olivier.Le_Meur/`.

[6] Face Lab 5, "Product," `http://www.seeingmachines.com/product/facelab/`.

[7] J.M. Henderson, "Regarding scenes," *Current Directions in Psychological Science*, vol. 16, pp. 219–222, 2007.

[8] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, pp. 643–659, 2005.

[9] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, pp. 1–16, 2012.

[10] R. Carmi and L. Itti, "Causal saliency effects during natural vision," in *Proc. ACM Eye Tracking Research and Applications*, Mar 2006, pp. 11–18.

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[12] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2169–2178.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 886–893.

[14] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[15] N. Riche, M. Mancas, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication,*, vol. 28, pp. 642–658, 2013.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.