Investigation on the Influence of Visual Attention on Image Memorability

Wulin Wang¹, Jiande Sun^{1,2(✉)}, Jing Li^{3,4}, Qiang Wu¹, and Ju Liu^{1,2}

¹ School of Information Science and Engineering, Shandong University, Jinan 250100, China
{jd_sun, wuqiang, juliu}@sdu.edu.cn, wangwulin@hotmail.com

² The Hisense State Key Laboratory of Digital-Media Technology, Qingdao 266061, China

³ School of Mechanical and Electrical Engineering, Shandong Management University,
Jinan 250100, China

lijingjdsun@hotmail.com

⁴ School of Information Science and Engineering, Shandong Normal University, Jinan 250014,
Shandong, China

Abstract. The research of image memorability has received increasing attention recently. In this paper, the influence of the visual attention based features on image memorability is explored, which is different from most of the existing studies focusing on various appearance features. In this paper, the dataset used by Isola et al. are adopted. The visual saliency map of each image in the dataset is generated via the visual attention model. The corresponding object-saliency map is obtained by replacing each object with its average visual saliency. The global, local, and joint spatial histograms based on the object-saliency map are obtained and the relationship between visual attention and memorability is explored based on these visual attention based features. The experiments are carried out by using two existing visual attention models and demonstrate that these mentioned visual attention based features are more effective than the appearance features to predict the image memorability.

Keywords: Image memorability · Visual attention · Scene understanding · Support vector machine

1 Introduction

Recently the research of image memorability has attracted increasing interests. If the features that are able to predict the memorability of images can be figured out, it is possible to modify the memorability of an image, which is promising in lots of image-related applications.

Image memorability is considered as the probability of correctly being detected a repetition of the image. Isola et al. showed that memorability is a trait intrinsic to images across different viewers [5]. The main issue of measuring the image memorability is to find the features that can predict the memorability of images accurately. There have been lots of studies on image memorability. Argembeau et al. took happy and angry expressions of face images into account and studied the effects of emotion on the memory [1].

Bainbridge et al. exploited more features than facial expressions in the study of the memorability of face images [2]. Isola et al. investigated the memorability of generic images, which include ordinary people or scenery and proposed to predict the image memorability according to different features, labels and attributes [3–5]. Khosla et al. considered the memorability depended on the difference between the initial image representation and its internal degraded version [6]. They predicted the memorability via using a noisy memory process of encoding images in the memory. Kim et al. showed that Weighted Object Area (WOA) and Relative Area Rank (RAR) can predict the image memorability [7]. Though these above researches have shown exciting performance, they still have some limitations. Only the appearance features are utilized to characterize the images and no visual factors are considered.

In the procedure of perceptual cognition, appearance features are on the first step and they stimulate the vision. The appearance features do not connect to memory directly, though they are the main features used in image analysis and play an important role in image understanding. Visual attention is a mechanism of visual system to deal with the regions with different visual saliences selectively and it is the step connected to the memory directly. The visual attention features can transfer the information of visual stimuli to the memory. Therefore, it is reasonable that visual attention has positive influence on image memorability.

In this paper, an investigation on visual attention based image memorability prediction is carried out, in which the features based on visual attention are utilized to predict the memorability. Two existing visual attention models are used to demonstrate the effectiveness of visual attention on the image memorability. The experiments demonstrate that these visual attention based features are more effective than the appearance features in predicting image memorability. Section 2 introduces the algorithm proposed by Isola et al. in [5]. Two visual attention models used in this paper are introduced in Sect. 3. How to investigate the influence of visual attention on image memorability is described in Section. Section 5 shows the experiment results and Sect. 6 gives the conclusion.

2 Previous Work on Image Memorability

In 2013, Isola et al. measured the memorability scores of 2400 images including people and natural scenery via a Visual Memory Game in [5]. These images are randomly sampled from different scene categories of the SUN dataset [10]. The memorability score of an image is defined as the probability of participants correctly detecting a repetition of the image in their study and this score is treated as the ground truth. Isola et al. trained a support vector regression SVR to map features of images to memorability scores. One half of the images scored by one half of the participants are used as the training examples, and the remaining are used as the test examples. It is data-dependent. During training, grid search is performed to choose the optimal parameters for SVR.

The performance is usually quantified by the Spearman's rank correlation (ρ). ρ is used to measure the statistical correlation between two groups of variables, and its theoretical value is between -1 and 1 . ρ is greater than zero means a monotonically increasing trend

of the two, and ρ is less than zero means a monotonically decreasing trend. Here ρ is used to evaluate the correlation between the feature and the ground truth memorability score of a photo. If ρ is greater than zero and it is large, it indicates this feature is advisable to representing the image memorability.

Many possible traits about image memorability were investigated in [5], which included external factors, various image-based, object-based and semantic features of images. External factors included different observers, time delays, context and subjective judgments about whether the image was memorable. Results demonstrated that memorability is an intrinsic and stable trait of an image.

For the image-based features, the hue, intensity and saturation of an image are investigated. The value of ρ between each of them and memorability was close to zero, which implied that each of them was correlated weakly with memorability. The object-based features such as labelled object counts, labelled object areas and spatial histograms of object distribution are also investigated by the means of machine learning. The experimental results showed that object based features had positive effect on image memorability. The semantic features included scene category and other semantic attributes labelled by human users, i.e. spatial layout of the scene, location of famous place and appearance of people (e.g., clothing, race, gender, etc.). When SVRs were trained to map these attributes to memorability scores, the performance of ρ was large. It implied that semantic attributes were efficient to characterize the memorability of an image. Meanwhile, they studied various global and local features algorithmically extracted from an image, i.e. SIFT, HOG, GIST and so on.

The study of Isola et al. focuses on the influence of appearance features on image memorability, which represents the image from the view of visual stimuli. However, what affects the memory directly is the visual reaction to visual stimuli, that is, visual attention. Therefore, we try to investigate the influence of visual attention on image memorability from the point of visual reaction.

3 Visual Attention Models

3.1 Itti's Attention Model

Visual attention has been proved to play an important role in the fields of image content analysis and understanding. Visual attention modelling simulates the behavior of human visual system by automatically producing saliency map of the target image and then detects out the regions of interests from the image, which are attractive to viewers.

Itti et al. proposed the most classical model which combined three kinds of low level appearance features, i.e. color, intensity and orientation, to extract salient regions [8]. The contrast of pixel was defined as its saliency, so the saliency map was constructed by computing the contrast of a pixel to its surroundings. Usually the regions with strong contrast to their surroundings had a high degree of saliency, which attracted people's attention, while the ones with weak contrast tended to be ignored. It was proved that this approach could extract approximate salient regions.

3.2 Attention Model with Cross-Layer Fusion

The attention model with cross-layer fusion was proposed by Sun et al. [9]. It was an improvement for the classical Itti's attention model by combining the local and global saliency. It computed the contrast saliency of the global and local layers respectively and combined the two layer saliency to generate a weight model and then optimized the global saliency using the weight model as a feedback from local layer. The final visual saliency map was obtained by performing the morphological post-processing. Figure 1 shows the framework of this visual saliency model. It could detect out more accurate salient regions.

Fig. 1. The framework of the cross-layer visual attention model. The optimization on the weighting cross the global and local layers improves the performance of attention model on detect out the body and boundary of the salient regions.

4 Image Memorability Based on Visual Attention

In this paper, the global, local and joint spatial histograms based on visual attention are investigated to predict image memorability. Figure 2 shows the framework. All objects in the images including people and scenery are labelled. The visual saliency map of each image is obtained via two existing visual attention models. The object regions are replaced with the average visual saliency in it, and the corresponding object-saliency map is generated. The global, local and joint spatial histograms are calculated based on the generated map.

Fig. 2. The framework of image memorability prediction based on visual attention. Visual saliency provides the visual weights to emphasize the relationship between the vision and memory.

To investigate the role of these spatial histograms based on visual attention in predicting image memorability, a support vector regression is trained to map these features to memory scores in this paper. The ground truth memorability scores are

regarded as labels. The experiment on 25 regression trials is performed and for each trial, both the images and participants are separated to two independent, random halves (same as [5]). Half of the images are sampled randomly from the 2222 images and their features are seen as training examples. Meanwhile, ground truth memorability scores corresponding to these examples are treated as training labels. The rest examples are used as test ones. During training, cross validation is performed to choose the optimal parameters for each SVR.

5 Experiments and Analysis

The dataset in [5], which consists of 2222 images selected from the SUN dataset in [10] are used in the experiments. The images are fully annotated with segmented object regions and are all resized to 256*256. A memorability score for each image is provided as the ground truth. Figure 3 shows three examples of experimental images. Figure 3(a) shows original images, and the visual saliency maps obtained by cross-layer based visual attention model are shown in Fig. 3(b). Figure 3(c) shows the object labelled maps, and Fig. 3(d) shows the object-saliency maps. It can be seen from Fig. 3 that the salient regions are captured well.

Fig. 3. (a) Example images, (b) visual saliency map, (c) objects labelled map and (d) object-saliency map. The difference between the visual and object saliency implies the influence of visual attention to memorability.

The performance is evaluated via Spearman's rank correlation (ρ). ρ is used to evaluate the correlation between the feature and the ground truth memorability score of an image. That ρ is greater than zero indicates the feature is advisable to repre-

senting the image memorability. The human consistency is set as $\rho = 0.75$ according to the study of [5], which is used as the upper bound in the performance of automatic methods, i.e., the ideal value. We evaluate the performance on 25 different training or testing splits of the data (the same splits as [5]) with an equal number of images for training and testing (i.e., 1111). The train splits are scored by one half of the participants and the test splits are scored by the other half of the participants. In addition, the performance can also be evaluated by average ground truth memorability over the images with the top N highest predicted memorability scores. The numerical comparison is listed in Table 1.

Table 1. Comparison between predictions and ground truth according to object based features, semantic features and SHVFs based on two attention models.

	T-20	T-100	B-20	ρ
Labelled object counts	82.4 %	79.0 %	55.2 %	0.44
Labelled object areas	84.3 %	82.3 %	53.5 %	0.48
Spatial histograms	85.1 %	82.3 %	52.7 %	0.48
Scene category	81.1 %	77.6 %	56.3 %	0.37
Combination of all global and local features	83.4 %	80.7 %	53.9 %	0.46
Spatial features in [7] (WOA)	84.9 %	82.2 %	51.2 %	0.49
Spatial features in [7] (RAR)	85.5 %	81.9 %	52.0 %	0.50
SHVF by Itti's attention model (global)	82.7 %	80.3 %	55.7 %	0.42
SHVF by Itti's attention model (local)	85.6 %	82.0 %	54.7 %	0.48
SHVF by Itti's attention model (joint)	85.9 %	82.3 %	54.5 %	0.48
SHVF by cross-layer attention model (global)	83.3 %	80.4 %	55.9 %	0.43
SHVF by cross-layer attention model (local)	85.5 %	82.3 %	54.1 %	0.49
SHVF by cross-layer attention model (joint)	85.7 %	82.6 %	53.1 %	0.49
Ideal value	86.9 %	84.3 %	39.6 %	0.75

Table 1 lists the comparisons on predictions and ground truth, where the predictions are obtained based on object based features, semantic features, combination of global and local features, spatial features in [7] and spatial histograms of visual features (SHVF) obtained by attention models respectively. The object-based features include labelled object counts, labelled object areas and spatial histograms. The semantic features refer to scene category. **T-20** and **T-100** denote the average ground truth memorability over the images with the top 20 and top 100 highest predicted memorability scores. **B-20** denotes the average ground truth memorability over the images with the top 20 lowest predicted memorability scores. In general, the value of Top 20 is about 81 %–85 % and ρ is 0.37–0.48 in [5].

In Table 1, T-20 obtained, by global, local and joint SHVF by two attention models is 85.9 %. It is 2.5 % higher than that obtained by the combination of all global and local features, 0.8 % higher than that obtained by object based features, i.e., spatial histograms and 4.8 % higher than that obtained by semantic features, i.e., scene category. T-100 obtained by global, local and joint SHVF by two attention models is 82.6 %. It is 1.9 %, 0.3 % and 5.0 % higher than that obtained by the combination of all global and local features, object based features and semantic features respectively. Moreover, local and joint SHVF are much closer to the ideal value than global SHVF and the other features. In addition, its highest Spearman's rank correlation ρ can reach 0.49, which is 0.03, 0.01 and 0.12 higher than the combination of all global and local features, object based features and semantic features and almost equal to that in [7]. It shows that the global, local and joint SHVF obtained by attention model are more effective in predicting image memorability. Especially the local SHVF is a little better than the global one due to the selective visual attention.

Figure 4 shows the comparison on ground truth and the predictions among labelled object counts, labelled object areas, spatial histograms, scene category and SHVFs by different attention models. The predicted memorability scores are sorted in descending order, which correspond to the sequence number 1–1111. In order to show the comparison clearly, only the top 100 numbers are shown in Fig. 4. The vertical axis corresponds to average ground truth memorability over the images with the top N highest predicted memorability scores. For instance, when N is equal to 100, the vertical axis corresponds to the average ground truth memorability over the images with the top 100 highest predicted memorability scores. From Fig. 4, it can be seen that when N is less than 100, the average ground truth memorability obtained by SHVF based on two attention models is higher than the others and much closer to the ideal value. It shows that visual attention based features can achieve better performance.

Figure 5 demonstrates the comparison on ground truth and the predictions between the combination of all global and local features and SHVF based on Itti's attention model. Compared to the combination of all global and local features, the performance achieved by local and joint SHVF by Itti's attention model are better. Especially when N is less than 20, the predicted memorability score is much closer to the ideal value. Although performance achieved by global spatial histograms are a little worse than those obtained by the combination of all global and local features, it can still imply that the features can be used to represent image. While the local and joint SHVF by Itti's attention model are proved to be more effective for representing image memorability.

Fig. 4. Comparison between ground truth and predictions according to object based features, semantic features and SHVF based on two attention models. The prediction performance considering attention models is better than those without considering attention models.

Fig. 5. Comparison between ground truth and predictions according to all global and local features and SHVF by Itti's attention model. The joint feature shows the reasonable better performance.

Figure 6 shows the comparison on the predictions and ground truth between the combination of all global and local features and SHVF based on cross-layer attention model. Figure 6 shows the similar comparison results as Fig. 5. It proves that the features based on visual attention are effective for representing image memorability.

Fig. 6. Comparison between ground truth and the predictions according to all global and local features and SHVF by cross-layer based model. The joint feature shows the reasonable better performance.

6 Conclusions

In this paper, we investigate the influence of the features based on visual attention on image memorability via the spatial histograms of visual features (SHVF) obtained by two visual attention models. The comparison results show that the visual attention has positive influence on predicting image memorability. Moreover, local and joint visual features have better performance than global ones, which is a reasonable reflection of the important role of selective attention in visual system. Future work will investigate the relationship between image memorability and other visual attention related-measures such as spatial and temporal distribution of visual attention regions, the order and intensity of visual attention without the labels.

Acknowledgments. The work is supported by the Natural Science Foundation of Shandong Province (ZR2014FM012). The contact author is Jiande Sun (jd_sun@sdu.edu.cn).

References

1. D'Argembeau, A., Van der Linden, M., Comblain, C., Etienne, A.-M.: The effects of happy and angry expressions on identity and expression memory for unfamiliar faces. *Cogn. Emot.* **17**(4), 609–622 (2003)
2. Bainbridge, W., Isola, P., Blank, I., Oliva, A.: Establishing a database for studying human face photograph memory. In: *Proceedings of the Cognitive Science Society* (2012)

3. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 145–152 (2011)
4. Isola, P., Parikh, D., Torralba, A., Oliva, A.: Understanding the intrinsic memorability of images. In: Advances in Neural Information Processing Systems (NIPS) (2011)
5. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes a photograph memorable. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(7), 1469–1482 (2013)
6. Khosla, A., Xiao, J., Torralba, A., Oliva, A.: Memorability of image regions. In: Advances in Neural Information Processing Systems (NIPS) (2012)
7. Kim, J., Yoon, S., Pavlovic, V.: Relative spatial features for image memorability. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 761–764 (2013)
8. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
9. Sun, J.D., Xie, J.C., Liu, J., Thomas, S.: Image adaptation and dynamic browsing based on two-layer saliency combination. *IEEE Trans. Broadcast.* **59**(4), 602–613 (2013)
10. Xiao, J., Hayes, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492 (2010)