Is Combinational Strategy Better For Image Memorability Prediction

Wenting Zhu

Abstract. Nowadays, multimedia developed so quickly that we are overwhelmed by all sorts of images in our daily life, including the images on the billboard, internet websites, magazines as well as TV shows. However, some images are easily forgotten while others remain stable memory in people's mind. As has been proved, images have inner characteristics that decide how memorable they are. So far, many different methods have been proposed to predict the memorability of an image. In this paper, we introduce some combinational strategies to improve the performance for image memorability prediction. It aims to answer whether combinational strategies can obtain better performance for image memorability. Experimental results show that the combinational methods indeed outperform unitary methods with traditional features, but have lower performance than that of CNN-based methods.

Keywords: image memorability, consensus-based strategy

1 INTRODUCTION

Whether voluntarily or forcedly, people tend to see a variety of images every day. However, we could not remember all of them although we human have a strong memorability. Actually, we can have a stable memorability of some images but easily forget others. Then what makes some images more memorable than others? Khosla reveals that there is a strong correlation between the images inner characteristic and their memorability. As the result shows, images with people, salient actions, events and central objects are generally more memorable to all of us than images with natural landscapes [1]. Figure 1 gives an example of images with different memorability. Hence, based on the correlation we find, it is possible that we can efficiently predict image memorability.

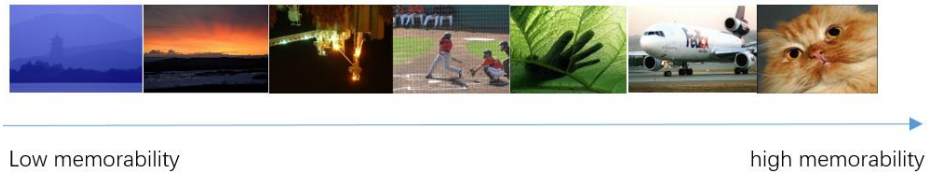

Fig. 1. Photos with different memorability in the dataset are shown, from left to right, the image memorability increase relatively. We can see from it that, in general, photos with salient objects are more memorable than photos with nature landscapes.

1.1 Related Work

By now, various methods to predict image memorability have been proposed. Some of them focus on effective feature extraction; while some of them focus on prediction model construction. We briefly review the existing work from the two categories as following.

For the feature selection, there are five kinds of features that have been used for image memorability prediction, including: simple image features [1], computer vision global features [1] [7], object statistics [1], object and scene semantics [1][2] and deep network features. Simple image features include mean value of three channel of color, hue, saturation, value and intensity in basic pixel statistics, which has been proved having week correlation with the memorability. The Computer vision global features consist of features that are related to the images gradient, texture, shapes and so on, such as Generalized Search Trees (GIST) [19], Histogram of Oriented Gradients (HOG2$\times$2) [17] and Scale-Invariant Feature Transform(SIFT) [21], that have stronger correlation with memorability scores. The object statistics represents non-sematic object counts, ranging from number of objects emerging in an image, mean pixel coverage over salient objects to max pixel coverage over salient objects, which does not perform well in the prediction [1]. Object and scene semantic captures many semantic attributes, for example, the spatial layout, aesthetic scores, emotion classes, locations and actions of the images, thus have a strong correlation with the image memorability scores. The last one is features from Convolutional Neural Networks (CNN), they are extracted from each layer of convolutional network, the higher the layer, the more effective the features.

For prediction models, there are three ways to achieve it – Support Vector Regression (SVR) [1][2], Linear Regression [3] and CNN [2]. As for the SVR (Support Vector Regression), it uses linear or RBF (Radial Basis Function) kernel to map selected feature values to a memorability score, with the selected attributes providing complementary information with each other. Linear Regression predicts by constructing a suitable loss function as well as searching for optimal parameters, aiming to get a minimum deviation with the ground truth scores. CNN is the newest and the most efficient way to predict, it uses a pre-trained Hybrid-CNN to train images from various splits and take the output layer of CNN, which are single real numbers, as the memorability scores.

Although fore-mentioned work attain relatively good performance, they belong to unitary methods (i.e. traditional, non-combinational, single-strategy methods) that try to predict the image memorability scores using a single model with a single kind of features. The memorability of images, serving as a complex inner characteristic, is difficult to be predicted by one typical attribute, either a complex feature or a simple images feature. Using one feature or simply putting several features into one attribute could lose important information, thus weaken the efficiency of estimation.

1.2 Our Work

To avoid the limitations, we introduce some combinational strategies for fusing these results from different unitary methods to obtain a better result for each image.

To this end, this paper should answer two questions:

- *How to combine the results from unitary methods?* To answer this question, many different combinational strategies, including both unsupervised and supervised models, are used in this paper.
- *Is combinational strategy useful for performance improvement for image memorability?* To validate the effectiveness of the proposed combinational strategies, we test then on a large scale of images. More analysis about experimental results is also given out.

The remainder of this paper is organized as follows. Section 2 introduces the process of feature selection and combining methods. Section 3 gives information about experimental setup and analysis of results. Section 4 discusses the overall experiment and give a conclusion.

2 COMBINATIONAL STRATEGIES FOR IMAGE MEMORABILITY

In this paper, we use ensemble learning to predict the image memorability, trying to attain a better performance by combining several unitary methods.

2.1 Framework

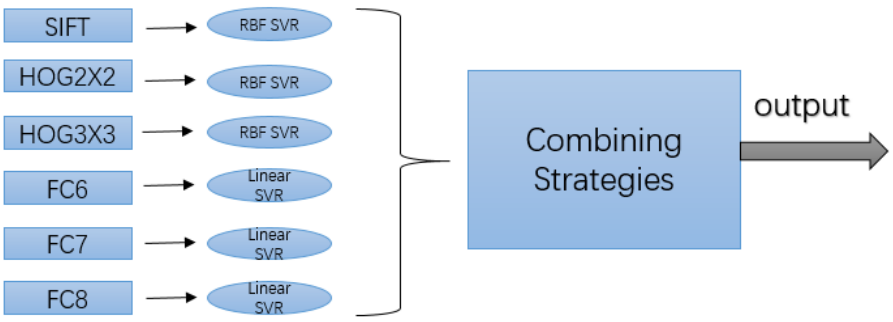

Fig. 2. The framework of the combinational strategy for image memorability

Fig.2 illustrate the framework of the proposed combinational strategy for the image memorability. It involves two key components: unitary method selection and combinational strategy design. For an image, each unitary method predicts its memorability score independently. Then all the scores are fused as a final score by a pro-

posed combinational strategy. Consequently, the two steps determine the final performance of the proposed combinational method.

2.2 Unitary Method Selection

Based on previous work [1][2][3], we can find that different characteristics of images have relatively different correlation with the memorability. In this paper, we select 6 characteristics that have best performances in the previous researches [1][2][3]. Three of them use traditional features (i.e. shallow features). The other three ones use deep features that generated by CNN [2]. The details of each unitary method with different features are discussed as follows:

- SIFT+SVR(RBF): The SIFT is one of the most frequently used in the region detector field [22]. It was first put forward by Lowe and it is now currently implied in various application fields. It is a region descriptor which extracts image features that are invariant to image scaling and rotation and partially invariant to changes in illumination and the 3D camera viewpoint. We use it as a training feature with 2100 dimensions for each image and train it with a RBF epsilon-SVR, for the reason that RBF kernel can provide a better result for the training process.
- HOG2×2+SVR(RBF): The HOG is extracted by computing and counting histogram of oriented gradients in local region of an image. Its combination with SVM have been widely used in pattern recognition, especially in pedestrian detection. Hence, we train this 2100 dimensional feature using a RBF epsilon-SVR to gain the results.
- HOG3×3+SVR(RBF): The HOG3X3 is similar to HOG2×2 in the extracting method, whereas the number of cells in each block is transformed from 2×2 to 3×3. We still used a RBF epsilon-SVR to train this feature to get the result.
- FC6,FC7, FC8+SVR(Linear): The FC6, FC7, FC8 features are all extracted from a hybrid fine-tuned CNN LaMem [2]. CNN is a multi-layer structure with several two-dimensional panel each layer and several independent neuron each panel. Each layer can be obtained by dealing with features of the previous layer. The FC6 is extracted from the 6^{th} layer of the CNN and FC7 is extracted from the 8^{th} layer, as for the FC8, it is the output of the CNN. We separately train these three features using Linear epsilon-SVR for the reason that it can attain a similar performance compared with RBF but with much less time cost.

Table 1. Selected features and their corresponding performance in memorability estimation.

Feature	SIFT	HOG2×2	HOG3×3	FC6	FC7	FC8
Training Method	RBF Kernel	RBF Kernel	RBF Kernel	Linear Kernel	Linear Kernel	Linear Kernel
ρ	0.48	0.46	0.45	0.60	0.64	0.64

2.3 Combinational Strategy

There are 6 kinds of classic combinational strategies considering ensemble learning, as [6] shows, these combining methods are based on the Euclidean distance selection among different results. The prior 5 methods are unsupervised methods while the last one is supervised.

Let us assume that we have n basis learning machine, dis represents the Euclidean distance between two numbers. p_i represents the predicted labels of each learning machine, we are dealing with the regression issue. Hence, the predicted labels are real numbers from 0 to 1, representing the memorability score of each image. These 6 combining strategies are as follows:

- Mean：mean value of all the basis learning machines result.

$$P_{mean} = \frac{\sum_{i=1}^{n} p_i}{n}$$

- Nearest2：mean value of the two closest Euclidean distance results.

$$P_{nearest2} = \frac{p_i + p_j}{2}; dis(p_i, p_j) = \min\{dis\}$$

- Nearest-N: mean value of the results with relative distances below (100+N) % of the minimum Euclidean distance.

$$P_{nearestN} = \frac{\sum p_m}{m}; dis(p_i, p_j) \le \frac{100+N}{100} \times \min\{dis\}$$

- No-N-Max：calculate Euclidean distance from each estimate to all the others, then subtract N estimates with the highest distance. Mean value of the left estimates. pi is the estimates with relatively lowest distance from all the others.

$$P_{NoNMax} = \frac{\sum_{i=1}^{n-N} p_i}{n-N}$$

- Median: extraction of the N results with the lowest distance from all the others. pi is the estimates with relatively lowest distance from all the others.

$$P_{Median} = \frac{\sum_{i=1}^{N} p_i}{N}$$

- Support Vector Regression Based Combination(SVRC): Support vector regression is a general method that estimate a continuous-valued function. It can also be used as a combinational strategy, which will be refer to as SVRC. The input of the SVRC is the results from several unitary methods. The input will be served as a new feature and we use SVR to train this new feature again. The output of this SVR is the combinational results. Through this method, we can obtain a better result for the reason that the combination process is supervised and can reduce the error tactfully.

3 RESULTS

3.1 Data Set and Experimental Setup

We use the LaMem dataset collected by Kholsa [6] in the following experiments. The dataset contains 58741 images with their memorability scores, including 45000 images for training, 10000 images for testing and 3741 images used for adjusting the parameter. The images are randomly divided for 5 times [3] and here we just randomly choose one split result to do our research. The dataset contains images of all kinds, ranging from landscape, human, object to geometrical photos, imitating various images people may see every day.

We use the newest toolbox libsvm-3.21 from Lin Chih-Jen [23] to train the SVRC method. We all use ρ, the spearman correlation between prediction and ground truth, as the estimation of the memorability prediction. When$\rho$is higher, the prediction is more precise.

3.2 Results of Combinational Methods

Table 2. Comparison among different combining strategies and human performance.

	Top20	Top100	Bottom20	Bottom100	ρ
Mean	90%	88%	53%	56%	0.62
Median	91%	90%	51%	55%	0.60
N2	92%	90%	50%	54%	0.54
N-100%	92%	90%	51%	54%	0.54
N-300%	91%	89%	51%	55%	0.56
No-1-Max	92%	89%	52%	55%	0.59
No-3-Max	91%	90%	52%	55%	0.58
SVRC	89%	88%	50%	53%	0.64
Human	86%	84%	47%	40%	0.75

In the column Top 20, we represent the mean value of memorability prediction for 20 highest memorable images. The column Top100 represents the mean value of memorability prediction for 100 highest memorable images. Similarly, Bottom 20 column represents the mean value of memorability prediction for 20 lowest memora-

ble images and Bottom 100 column represents the mean value of memorability prediction for 20 lowest memorable images. The first column illustrates 8 different strategies and human performance, which serves as ground truth.

The results of the proposed combinational methods on the LaMem set are listed in Table 2. As we can see from the table above, these eight combining methods have a performance ranging from 0.53 to 0.64, which indicates that the different combinational strategies results in different performance. Furthermore, among those eight combining methods, supervised combinational method out-performsthe others. As for the other six unsupervised methods, mean method gets the best performance with ρ=0.62.

3.3 Unitary Methods V.S. Combinational Methods

Table 3. Performance of unitary methods compared with combinational strategies.

		Top 20	Top 100	Bottom 20	Bottom 100	ρ	Mean
Unitary Method	SIFT	94%	91%	50%	54%	0.48	0.56
	HOG2×2	93%	90%	50%	54%	0.46	
	HOG3×3	92%	90%	50%	54%	0.45	
	FC6	83%	82%	59%	61%	0.60	
	FC7	103%	101%	41%	45%	0.64	
	FC8	94%	93%	48%	51%	0.64	
Combinational Mehods	Mean	90%	88%	53%	56%	0.62	0.58
	Median	91%	90%	51%	55%	0.60	
	N2	92%	90%	50%	54%	0.54	
	N-100%	92%	90%	51%	54%	0.54	
	N-300%	91%	89%	51%	55%	0.56	
	No-1-Max	92%	89%	52%	55%	0.59	
	No-3-Max	91%	90%	52%	55%	0.58	
	SVRC	89%	88%	50%	53%	0.64	
	Human	86%	84%	47%	40%	0.75	0.75

To validate the performance improvement of combinational methods, we compare the performance between unitary methods and combinational methods, as shown Table 3. We can see that the combinational methods do not obtain obvious performance improvements. Although the performance of combinational methods is higher than these unitary methods with traditional features (SIFT, HOG2×2, HOG3×3) whose performance ranges from 0.45 to 0.48, it is still lower than 0.64, which is the performance of CNN-based methods.

Why these combinational strategies do not outperform the CNN-based method? One of the reasons we suppose the combinational strategy can improve the performance lies in that combinational method can balance various error of different unitary methods. Hence, an important premise is that the features should come from different models. However, fc6, fc7, fc8 are obtained from the same CNN model, thus after combination, the errors in these three features could not be balanced or reduced. In

addition, the other three traditional features have lower performance than features obtained from CNN. Accordingly, these combinational methods still could not attain a better overall performance.

3.4 Comparison of Combinational Method based on Different Unitary Methods

Table 4. Performance of combinational method on different unitary methods

	ρ	Mean	SVRC
SIFT	0.48		
HOG2×2	0.46	0.46	0.51
HOG3×3	0.45		
FC6	0.60		
FC7	0.64	0.63	0.64
FC8	0.64		

To compare the SVRC method based on different unitary methods. We compare the SVRC based on the unitary methodswith traditional features (SIFT, HOG2×2, HOG3×3) (denoted as SVRC_T)and the SVRC based on the unitray methods with deep features (CNN) (denoted as "SVRC_D"), as shown in Table 4. We can see that the combinational method SVRC_T attains a better performance as 0.51. Hence, when using the unitary methods, combination can improve the performance by balance the different kinds of errors. However, when we combine three CNN features (FC6, FC7, FC8) using SVRC-D, the performance improvement is very limited.

3.5 Analysis and Discussion

According to the above experimental results, we can conclude that the combinational strategies could not attain a better performance for the image memorability prediction.The main reason is that the CNN features are obtained from same model, reducing the efficiency of combinational strategies. Moreover, relatively low performance also weaken the efficiency of combinational strategies, leading the performance not better.

Next. Let us look at the images that were best predicted using different combining ways. Figure 3 exactly demonstrates top 6 images with best estimation using those methods. We can easily realize different methods are suitable for predicting different kinds of images. For example, Mean-methodhave these top 6 best predicting images all with the salient objects in the center and also most of them have human face. As for Median, it could perform better when predicting pictures with nature or daily scenes without human. The next five methods are different, they perform well on a variety of themes, thus they are suitable for predicting a dataset containing all kinds of images.

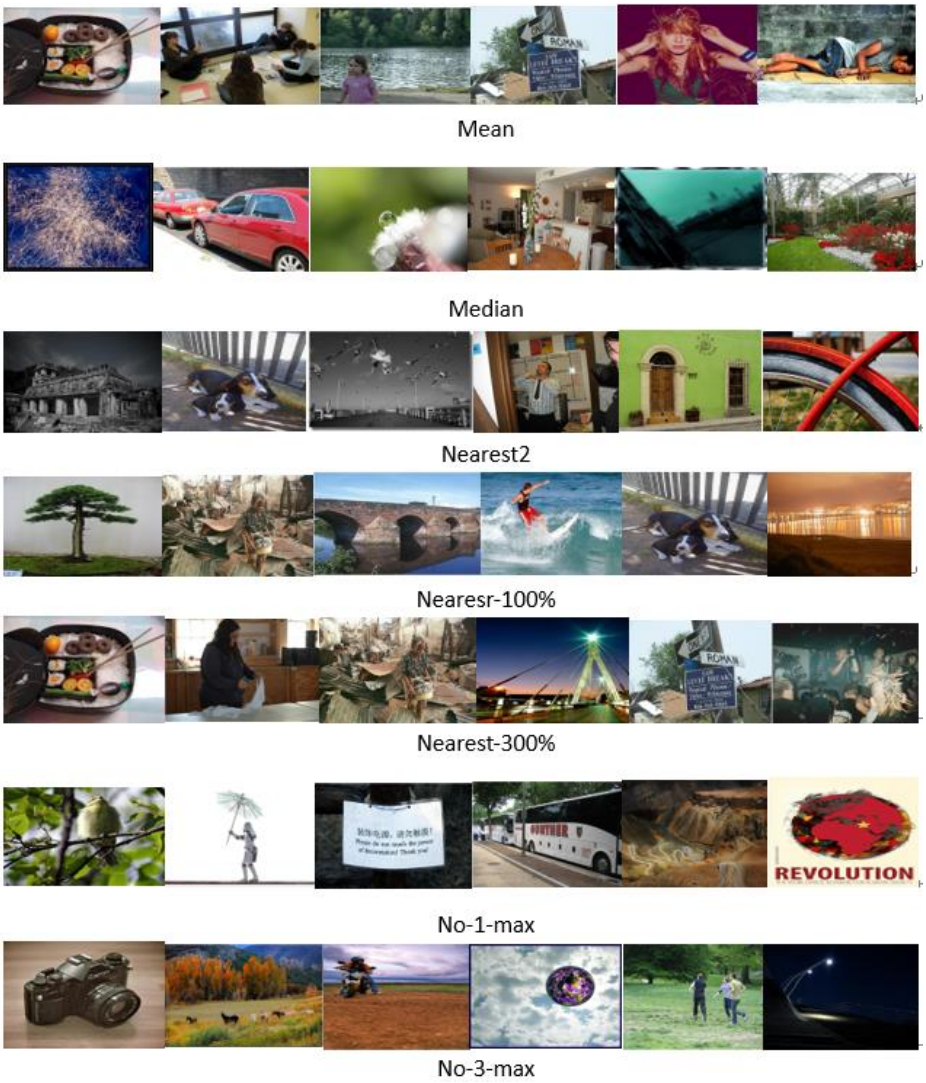

Fig. 3. This figure shows the top 6 photographs with the least estimation error of image memorability compared to the ground truth, in this figure, we can see that what kind of images are different methods suitable.

4 CONCLUSION

Image memorability prediction is a newly and interesting topic, with a galaxy of applications in our day-to-day life. For example, advertising agency could regard this as a tool to estimate the memorability of images, striving to leave a higher memorable impression on watchers' mind. Educational textbook could thus select more memora-

ble pictures to help students remember knowledge. Facebook fans can select a photo of themselves with higher memorability before publishing it. In this way, image memorability prediction is also a practical topic.

In this paper, we argue a new way to predict memorability of pictures by seeking for a combinational method, we obtain performance of 0.64 while still discuss difference and similarity of those 8 various combining methods as well as their under-lying reasons.

5 ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China (No. 61370038).,

6 REFERENCE

1. P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes a photograph memorable? IEEE TPAMI, 2014.
2. A. Khosla, A. S. Raju, A. Torralba and A. Oliva. Understanding and Predicting Image Memorability at a Large Scale. International Conference on Computer Vision (ICCV), 2015 DOI 10.1109/ICCV.2015.275
3. Houwen Peng, Kai Li, Bing Li, Haibin Ling, Weihua Xiong, Weiming Hu. Predicting Image Memorability by Multi-view Adaptive Regression. 2015 ACM. ISBN 978-1-4503-3459-4/15/10.
4. M. J. Huiskes and M. S. Lew. The MIR Flickr retrieval evaluation. In ACM MI.2008.
5. N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large scale database for aesthetic visual analysis. CVPR, 2012. 2, 5.
6. S. Bianco, F. Gasparini, and R. Schettini. A Consensus Based Framework for Illuminant Chromaticity Estimation. Journal of Electronic Imaging, April, 2008.
7. P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In NIPS, pages 2429{2437, 2011.
8. B. Celikkale, A. Erdem, and E. Erdem. Visual attention driven spatial pooling for image memorability. In CVPRW, pages 976{983, 2013.
9. E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In CVPR, 2007.
10. T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. Proc Natl Acad Sci, USA, 105(38), 2008.
11. H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. NIPS, 1997.
12. A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In CVPR, 2009.
13. A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva. Image memorability and visual inception. In SIGGRAPH Asia 2012 Technical Briefs, page 35. ACM, 2012.
14. A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In NIPS, 2012.
15. T. Konkle, T. F. Brady, G. A. Alvarez, and A. Oliva. Scene memory is more detailed than you think: the role of categories in visual long-term memory. Psych Science, 21(11), 2010.

16. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. IJCV, 2015.
17. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In CVPR, 2010.
18. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In NIPS, 2014.
19. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," Int. J. Comput. Vision, vol. 42, no. 3, pp. 145–175, May 2001.
20. E. Shechtman andM. Irani, "Matching local self-similarities across images and videos," in Proc. IEEE Conf. CVPR, Minneapolis, MN, USA, 2007.
21. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in Proc. IEEE Computer Soc. Conf. CVPR, Washington, DC, USA, 2006.
22. Lingua, A.; Marenchino, D.; Nex, F. Performance Analysis of the SIFT Operator for Automatic Feature Extraction and Matching in Photogrammetric Applications. *Sensors* 2009, *9*, 3745-3766.
23. Chih-Chung Chang, Chih-Jen Lin. A library for support vector machines. ACM Transacion on Intelligent System and Technology (TIST) archive, April 2011.