SCIENCE CHINA
Information Sciences

• Supplementary File •

Large Margin Deep Embedding for Aesthetic Image Classification

Guanjun GUO[1], Hanzi WANG[1*], Yan YAN[1], Liming ZHANG[2] & Bo LI[3]

[1]*Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen 361005, Fujian, P. R. China;*
[2]*Faculty of Science and Technology, University of Macau, Macau 999078, China;*
[3]*Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, P. R. China.*

1 Experiments

In this section, we evaluate the performance of the proposed LMDE method on two large-scale visual aesthetic datasets: the AVA dataset [1] and the CUHKPQ dataset [2]. The proposed method is implemented in Torch7 [3] and trained on a computer equipped with a GTX-Titan GPU by using the proposed fast training strategy, where γ, β and S are experimentally set to 0.2, 1.0 and 5.0, respectively. The experiments include two parts. The first part evaluates the performance of the proposed method for aesthetic image classification with different parameter settings. The second part evaluates the proposed method for aesthetic image classification and compares it with several state-of-the-art aesthetic image classification methods on the two popular datasets.

1.1 The Datasets

We use two datasets (i.e., the CHUKPQ dataset and the AVA dataset) to evaluate the competing aesthetic image classification methods. The CHUKPQ dataset contains about 30,000 images, which are collected from a variety of photography websites. Each image in this dataset is clearly labelled as either a low-quality image (with a low aesthetic score) or a high-quality image (with a high aesthetic score). There are 10,525 high-quality images and 19,167 low-quality images labelled in the CHUKPQ dataset.

The AVA dataset contains over 250,000 images collected from the website of Dpchallenge. Each image has about 210 aesthetic scores ranged from 1 to 10. We use 230,000 images for training and the remaining 20,000 images for test. We divide the training images into two categories (i.e., low-quality images and high-quality images) for training the proposed LMDE method. Following the strategy used in [1, 4], a parameter δ is used to discard the ambiguous images from the training set: The images with an average score smaller than $5-\delta$ are referred to the low-quality images. The images with an average score larger than or equal to $5+\delta$ are considered as the high-quality images. The images with an average score between $5-\delta$ and $5+\delta$ are considered as the ambiguous images, and these images are discarded. In the implementation, we experimentally set the value of δ to be 1, by which 52,207 high-quality images and 21,140 low-quality images are finally collected. To alleviate the class imbalance problem, we simply augment the low-quality images by flipping each low-quality image horizontally. Thus, we obtain 42,280 low-quality images in total. All the obtained images are split into 738 batches, and each batch consists of 128 images.

* Corresponding author (email: hanzi.wang@xmu.edu.cn)

Table 1 The structures of the two CNN nets used in the proposed LMDE method.

Layer	Layer1	Layer2	Layer3	Layer4	Layer5	Layer6	Layer7	Layer8
Net1	C(5, 3, 64)	P(2)	C(5, 64, 64)	P(2)	C(5, 64, 64)	P(2)	C(5, 64, 64)	P(2)
Net2	C(5, 3, 64)	P(2)	C(5, 64, 64)	P(2)	C(5, 64, 64)	P(2)	C(5, 64, 64)	SPP(2, 3, 4)

Table 2 The confusion matrices obtained the proposed LMDE method using two different CNN structures (i.e., Net1 (a) and Net2 (b)) on the CHUKPQ dataset.

	High-quality	Low-quality
High-quality	4,605	657
Low-quality	672	8,911

(a)

	High-quality	Low-quality
High-quality	5,026	236
Low-quality	536	9,047

(b)

1.2 The Influence of the Parameters

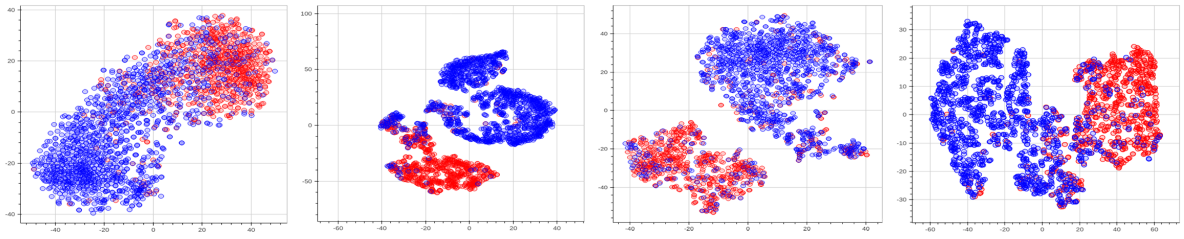
The proposed LMDE method contains several crucial parameters, which have influence on the performance of LMDE. In this subsection, we examine how these parameters affect the performance of the proposed LMDE method for aesthetic image classification.

1.2.1 *Using the SPP Layer*

A conventional preprocessing step for training a CNN model is to resize all images of a dataset into a fixed size. However, resizing a visual pleasing image may potentially damage its aesthetics. Therefore, we conduct experiments to evaluate whether the operation of resizing images into a fixed size will affect the classification performance of LMDE. We train two LMDE models with different CNN structures (called as Net1 and Net2) on the CHUKPQ dataset. The two CNN structures are respectively shown in Table 1. In Table 1, the convolutional layer is denoted by C(k, nIn, $nOut$), where nIn and $nOut$ respectively denote the number of input feature maps and the number of output feature maps, and the kernel size of filters is $k \times k$. Each convolutional layer is activated by a rectfied linear unit (ReLU) [5]. P(s) denotes the max-pooling layer, whose kernel size is $s \times s$. SPP(s_1, s_2, $\ldots$, s_z) denotes the spatial pyramid pooling layer, which partitions each input feature map into divisions from fine levels to coarse levels, and aggregates local features in the divisions. In total, a z-levels pyramid is generated and the zth level is partitioned into $s_z \times s_z$ divisions. The only difference between the two CNN nets in Table 1 is that in the last layer, Net2 uses the SPP layer, while Net1 uses a max-pooling layer. We report the classification results obtained by LMDE with the two CNN nets, which are given in the two confusion matrices in Table 2. As shown in Table 2, the number of high-quality images predicted correctly by Net2 is more than that by Net1. The main reason is that images are not resized into a fixed size in the process of training Net2, which uses the SPP layer in the last layer. Therefore, the SPP layer is beneficial to the proposed LMDE method for aesthetic image classification. In the following experiments, the CNN structure of Net2 in Table 1 is used if not specified. Rojas et al. [6] also observe that aesthetics can be affected after resizing images, and they avoid resizing input images by using fully convolutional nets as the classifier. However, [6] does not consider the problem of large intra-class variations in aesthetic images.

1.2.2 *Different Loss Functions*

We respectively set the value of λ in Eqn. (11) in the letter to 0, 1.0 and 0.5, and train the proposed LMDE method on the CHUKQP dataset, by which the proposed LMDE method is trained with the hinge loss function (i.e., λ=0), the triplet loss function (i.e., λ=1.0) and the joint loss function (i.e., λ=0.5), respectively. For simplicity, we call the LMDE method with different values of λ as LMDE$_H$, LMDE$_T$, LMDE$_J$, respectively. The CNN nets in LMDE$_H$, LMDE$_T$ and LMDE$_J$ are respectively used to extract features from 2,000 images randomly selected from the



(a) (b) (c) (d)

Figure 1 The visualization of the features obtained by the proposed LMDE method with different parameter settings (the red and blue points denote the features of high-quality images and the features of low-quality images, respectively). (a)-(c) The visualization of the features obtained by the LMDE method trained with the triplet loss function, the hinge loss function and the joint loss function, respectively. (d) The visualization of the features obtained by the CNN model, which is trained with the softmax loss function.

Table 3 The classification accuracy obtained by the proposed LMDE method with different parameter settings and the trained CNN Net2 model on the CHUKPQ dataset.

Loss Function	Accuracy
Hinge (i.e., LMDE$_H$)	83.20%
Triplet (i.e., LMDE$_T$)	85.31%
Softmax (i.e., CNN Net2)	89.15%
Joint (i.e., LMDE$_J$)	**94.80%**

CHUKQP dataset. Using the t-SNE visualization algorithm [7], the extracted features are respectively shown in Figure 1(a)-(c). As shown in Figure 1, the features extracted by LMDE$_J$ show the maximum inter-class separability. In addition, we also train Net2 with the standard softmax loss (called as CNN Net2) as a comparison. The features obtained by the trained Net2 are shown in Fig. 1(d), which shows better inter-class separability than those obtained by LMDE$_H$ and LMDE$_T$. However, the features obtained by CNN Net2 show less inter-class separability than those obtained by LMDE$_J$, since LMDE$_J$ explicitly encourages inter-class separability between the learned features from different classes. Table 3 reports the classification accuracy obtained by LMDE$_H$, LMDE$_T$, LMDE$_J$ and CNN Net2 on the CHUKPQ test dataset. As can be seen, LMDE$_J$ obtains the best accuracy, since the features obtained by LMDE$_J$ show the best inter-class separability. Table 4 reports the classification accuracy obtained by LMDE$_H$, LMDE$_T$, LMDE$_J$ and the trained CNN Net2 on the AVA test dataset. As can be seen, LMDE$_J$ still obtains the best accuracy. Generally, the results obtained by the proposed method with different parameter settings on the AVA dataset are worse than those on the CHUKPQ dataset. The main reason is that the images in the AVA dataset have more significant variations than those in the CHUKPQ dataset.

1.3 Comparison with the State-of-the-art Methods

In this subsection, we evaluate the performance of the proposed method for aesthetic image classification and compare it with several state-of-the-art methods on the two visual aesthetic datasets. We select five image aesthetic

Table 4 The classification accuracy obtained by the proposed LMDE method with different parameter settings and the trained CNN Net2 model on the AVA dataset.

Loss Function	Accuracy
Hinge (i.e., LMDE$_H$)	72.17%
Triplet (i.e., LMDE$_T$)	76.41%
Softmax (i.e., CNN Net2)	74.20%
Joint (i.e., LMDE$_J$)	**85.15%**

Table 5 The classification accuracy obtained by CAP, CB and the proposed LMDE method for aesthetic image classification on the CHUKPQ dataset.

Method	Animal	Plant	Static	Architecture	Landscape	Human	Night	Overall
CAP [8]	78.61%	76.38%	71.74%	73.86%	77.53%	76.94%	64.21%	77.92%
CB [9]	89.37%	91.82%	90.69%	**92.75%**	94.68%	**97.40%**	84.63%	92.09%
LMDE	**95.54%**	**96.70%**	**94.93%**	92.43%	**96.49%**	95.35%	**92.18%**	**94.80%**

Table 6 The classification accuracy obtained by the proposed LMDE method and the other three competing methods for aesthetic image classification on the AVA dataset.

Method	Accuracy
FV [1]	68.00%
RAPID [4]	73.70%
MPAN [10]	75.41%
LMDE	**85.15%**

classification methods for comparison: CAP [8], CB [9], FV [1], RAPID [4] and MPAN [10]. The first three methods are the representative methods based on hand-crafted features for aesthetic classification. The latter two methods are based on deep learning. Following [9, 10], we respectively compare the proposed method with CAP and CB on the CHUKPQ dataset, and compare it with the other three competing methods (i.e., FV, RAPID and MPAN) on the AVA dataset.

1.3.1 *Results on the CHUKPQ dataset.*

The CHKUPQ dataset consists of seven classes of images: *animal*, *plant*, *static*, *architecture*, *landscape*, *human* and *night*. Following the strategy of selecting training and test sets in [9], we randomly select half of the high-quality images and half of the low-quality images as the training set and keep the remaining images as the test dataset. The proposed method is trained on the training set, which takes about 36 hours.

Table 5 reports the classification accuracy obtained by the CAP, CB and LMDE methods. As can be seen, the overall average classification accuracy obtained by the proposed LMDE method is 94.8%, which is the highest among the three competing methods and it is higher than those of CAP and CB by 16.9% and 2.7%, respectively. The CAP method obtains the worst average classification accuracy since it only uses several regional features (including color saturation, texture feature, the depth of field feature, etc.), which are less effective for the task of aesthetic image classification. The CB method uses more effective features (including the regional features and the global features) than those used in the CAP method. Thus, the CB method obtains better average classification accuracy than the CAP method. Since the proposed LMDE method takes the advantage of CNN to directly learn the features, the proposed LMDE method obtains higher average classification accuracy than the CB method. More specifically, the proposed LMDE method respectively obtains 1-5% improvements on the classes of *animal*, *plant*, *static* and *landscape* compared with the CB method, and it also achieves a comparative result to CB on the class of *architecture*.

1.3.2 *Results on the AVA dataset.*

Table 6 lists the average classification accuracy obtained by the proposed LMDE method and the other three competing methods for aesthetic image classification on the AVA dataset. As we can see, the proposed LMDE method clearly outperforms the other three competing methods. The FV method [1] obtains the worst performance since it only uses the hand-crafted features, which are less effective for classification. To the best of our knowledge, the RAPID method is the first to use CNN for the task of aesthetic image classification. The powerful ability of

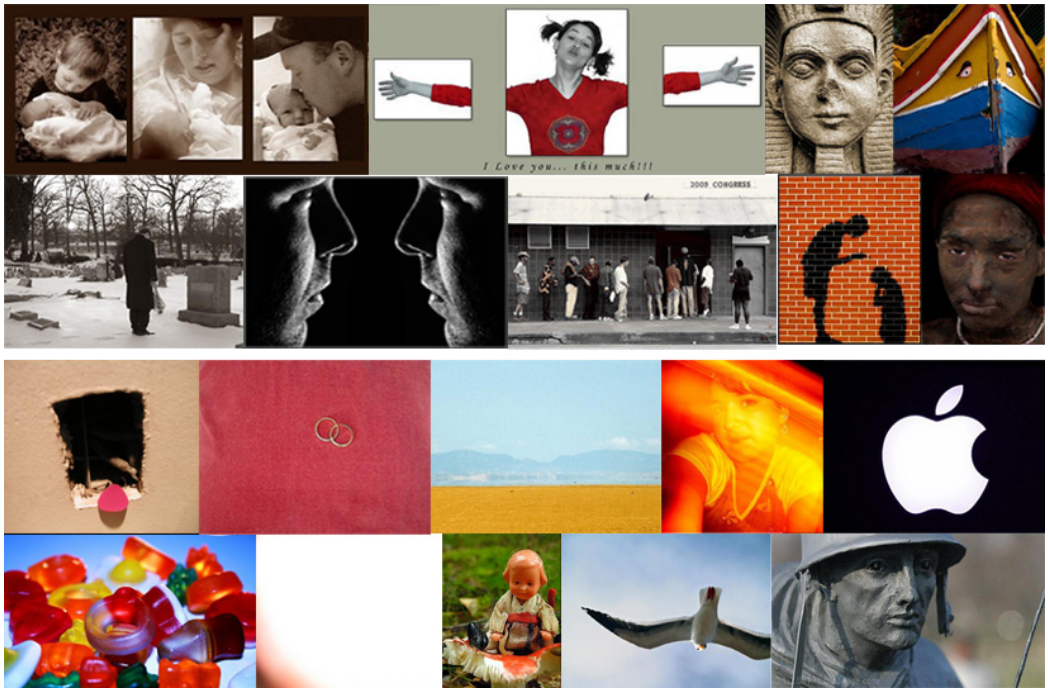


Figure 2 Failure examples obtained by the proposed method. The first two rows and the last two rows show false positive examples and false negative examples, respectively.

feature learning with CNN makes RAPID obtain better performance than the FV method. However, RAPID does not consider that the operation of resizing images may change the aesthetics of the images. In contrast, MPAN uses the SPP layer and aggregates fine-grained details from multiple patches. Thus, MPAN achieves about 2% gain over RAPID. The proposed LMDE method explicitly maps the images to a compact space, in which the features of aesthetically pleasing photos are nearby and they are far from those of aesthetically unpleasing photos. Moreover, the proposed LMDE method encourages inter-class separability between the features of the images from different classes with a large margin. Thus, the proposed method outperforms all the other competing methods by about 9%-17% in terms of classification accuracy.

1.3.3 *Limitation Analysis.*

We give several failure examples obtained by the proposed method in Figure 2. As shown in Fig. 2, the false positive images are those that have attractive stories or creative ideas. The current aesthetic classification methods cannot recognize high-level semantic information. Thus, the proposed LMDE method and other aesthetic classification methods usually classify them into low-quality images. In contrast, the false negative images usually have clear topics, vivid colors or simply aesthetic but they have trivial stories. Thus, photographers give the false negative images low scores, but the aesthetic classification methods usually classify them into high-quality images.

References

1 Murray N, Marchesotti L, Perronnin F. Ava: A large-scale database for aesthetic visual analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012. 2408–2415
2 Luo W, Wang X G, Tang X O. Content-based photo quality assessment. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2011. 2206–2213
3 Collobert R, Kavukcuoglu K, Farabet C. Torch7: A matlab-like environment for machine learning. In *Workshop on Proc. Adv. Neural Inf. Process. Syst.*, 2011. 1–6
4 Lu X, Lin Z, Jin H L, Yang J C, Wang J Z. Rapid: Rating pictorial aesthetics using deep learning. In *ACM Multimedia*, 2014. 457–466
5 George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, 2013. 8609–8613
6 Wang Z Y, Dolcos F, Beck D, Chang S Y, Huang T S. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint*, 2016, abs/1601.04155

7 Van L P, Maaten D, Hinton G E. Visualizing high-dimensional data using t-sne. *Journal of Mach. Learn. Research*, 2008, 9:2579–2605
8 Datta R, Joshi D, Li J, Wang J Z. Studying aesthetics in photographic images using a computational approach. In *Proc. Eur. Comput. Vis. Conf.*, 2006. 288–301
9 Tang X O, Luo W, Wang X G. Content-based photo quality assessment. *IEEE Trans. on Multimedia*, 2013, 15:1930–1943
10 Lu X, Lin Z, Shen X H, Mech R, Wang Z J. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2015. 990–998