UNDERSTANDING THE AESTHETIC STYLES OF SOCIAL IMAGES

Yihui Ma[1], *Jia Jia*[1], *Yufan Hou*[1], *Yaohua Bu*[12], *Wentao Han*[1*]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
[2]Academy of Arts & Design, Tsinghua University, Beijing, China

mayihui12@foxmail.com, jjia@mail.tsinghua.edu.cn, evan9669@126.com,
byh15@mails.tsinghua.edu.cn, hanwentao@tsinghua.edu.cn

ABSTRACT

Aesthetic perception is nearly the most direct impact people could receive from images. Recent research on image understanding is mainly focused on image analysis, recognition and classification, regardless of the aesthetic meanings embedded in images. In this paper, we systematically study the problem of understanding the aesthetic styles of social images. First, we build a two-dimensional Image Aesthetic Space (IAS) to describe image aesthetic styles quantitatively and universally. Then, we propose a Bimodal Deep Autoencoder with Cross Edges (BDA-CE) to deeply fuse the social image related features (i.e. images' visual features, tags' textual features). Connecting BDA-CE with a regression model, we are able to map the features to the IAS. The experimental results on the benchmark dataset we build with 120 thousand Flickr images show that our model outperforms (+5.5% in terms of MSE) alternative baselines. Furthermore, we conduct an interesting case study to demonstrate the advantages of our methods.

***Index Terms*—** Aesthetic style, social image, autoencoder, dimensional space

1. INTRODUCTION

Aesthetic perception can be regarded as images' most direct impact on people. Taking the paintings as an example, impressionistic works tend to be bright and warm, while ink wash paintings present a clear and cool style. Although many researchers have been dedicated to image analysis, recognition and classification [1, 2, 3], the aesthetics-oriented image understanding is still in the early stage. If appreciating image aesthetic styles can be achieved automatically, it will contribute to the development of many research fields (e.g. image recognition, image retrieval, etc.), and make image relevant applications more humanized.

In recent years, many efforts have been made towards understanding images. [4] uses deep residual learning framework to achieve high performance in ImageNet classification tasks. [5] releases ConvNet models to facilitate deep visual representations. These work performs well in image recognition and classification tasks, but they are not involved with images' aesthetics. Focused on aesthetic visual analysis, [6] introduces a large-scale database for image aesthetics. [7] describes an approach to predicting image styles on Flickr photographs and paintings datasets. However, these research defines the styles as a few categories, which are not enough to cover various images' styles. Focused on clothing styles, [8] proposes to appreciate the aesthetic styles of upper-body menswear, and [9] tries to better understand fashion styles of clothing collocations. Both of them try to express the aesthetic meaning of clothing images quantitatively, while image categories are far more than just clothing (e.g. landscapes, paintings, portraits, etc.). Thus, there still remain two tough challenges for us: 1) how to quantitatively describe the aesthetic styles of various images, 2) how to model the subtle relationship between image related features and aesthetic styles.

In this paper, we systematically study the problem of understanding the aesthetic styles of social images, and propose our solutions from two aspects for the above questions. First, we build a two-dimensional Image Aesthetic Space (IAS) based on the image-scale space proposed by Kobayashi [10]. We collect the most often used 684 aesthetic words from the world's largest image social network Flickr, and coordinate them in the space by computing their semantic distances from Kobayashi's keywords using WordNet::Similarity [11]. Second, we propose an aesthetics-oriented multimodal deep learning model, Bimodal Deep Autoencoder with Cross Edges (BDA-CE), to deeply fuse the social image related features (i.e. images' visual features, tags' textual features) and learn their high-level joint representations. Connecting BDA-CE to a regression model, we finally map the joint representations to the IAS. The experimental results on the benchmark dataset we build with 120 thousand Flickr images

* Corresponding Author

indicate that our model outperforms (+5.5% in terms of MSE) alternative baselines. Furthermore, we show an interesting case study to demonstrate the advantages of our methods.

2. METHODS

2.1. Image Aesthetic Space

For art design, Kobayashi [10] proposes 180 keywords in 16 categories and defines their coordinates in two-dimensional (warm-cool and hard-soft) image-scale space. However, this space is mainly designed for color combinations, and the keywords are basic adjectives that may not be suitable for describing image aesthetic styles. Thus, we collect user-labeled aesthetic words from the world's largest image social network Flickr, and build the Image Aesthetic Space based on Kobayashi's keywords.

Exploring image aesthetic words. In order to discover which aesthetic words are usually used by people to describe various social image styles, we first crawl the user-labeled words of Flickr images posted in the last three years. Using WordNet [12], we retain only adjectives. Next, we manually remove those not often used to describe image aesthetic styles (such as "American" or "red"). Finally, we get 684 aesthetic words representing the aesthetic styles of social images.

Building the Image Aesthetic Space. To determine the coordinates of these new aesthetic words, we calculate the semantic distances between Kobayashi's keywords and the new aesthetic words using WordNet::Similarity [11]. For a word to be coordinated, we choose three Kobayashi's keywords with the shortest semantic distances, the arithmetic mean of which can be regarded as the coordinates. In this way, we build the Image Aesthetic Space (IAS). In order to present it more clearly, we show a simplified illustration in Fig. 1. To enhance the readability of the space's visualization, we only retain 83 words that occur frequently on Flickr and adopt Kobayashi's 16 categories to present them separately.

2.2. Bimodal Deep Autoencoder with Cross Edges

Intuition. In our training task, we need to map the image and text features of social images to the IAS. Due to the very different representation of these features, we need to fuse the modalities. Thus, we propose an aesthetics-oriented multimodal deep learning model, named Bimodal Deep Autoencoder with Cross Edges (BDA-CE), to fulfill this task. After modality fusion, we can use a regression model to map the joint representation of two modalities to the IAS.

The structure of BDA-CE. Given an image $v_i \in V$, the initial input vector x_i^I represents the image features vector and x_i^T represents the text features vector. The training target is to map $\{x_i^I, x_i^T\}$ to a deep representation of the two modalities. The whole training process of our model is described as the following stages, which is also shown in Fig. 2.

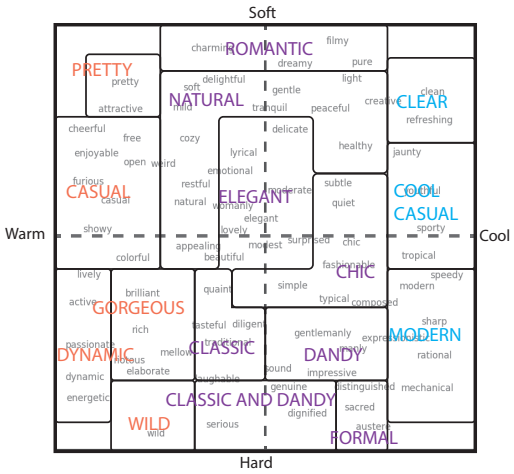


Fig. 1. The simplified illustration of IAS.

Stage 1: We train a deep autoencoder for each modality. Due to the similarity, we take the image modality as example and explain the network in detail. As shown in Fig. 2(a), the initial input vector x_i^I is reconstructed into $\hat{x}_i^I$. For both autoencoders, we adopt the same number of hidden layers to facilitate the training stages later. Formally, supposing the autoencoder has N_h hidden layers, the recursion formula between adjacent layers is defined as:

$$h_{I,i}^{(l+1)} = sigmoid(W_I^{(l)} h_{I,i}^{(l)} + b_I^{(l)}) \tag{1}$$

where $h_{I,i}^{(l)}$ denotes the vector of the lth hidden layer, $W_I^{(l)}$ and $b_I^{(l)}$ are the parameters between lth layer and $(l+1)$th layer and $sigmoid$ is the sigmoid function ($sigmoid(x) = \frac{1}{1+e^{-x}}$). Specially, $h_{I,i}^{(0)} = x_i^I$ and $h_{I,i}^{(N_h+1)} = \hat{x}_i^I$. The cost function to evaluate the difference between x_i^I and $\hat{x}_i^I$ is defined as:

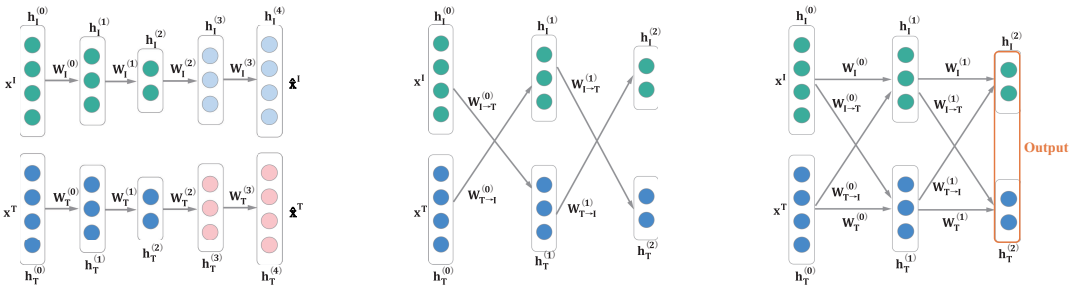
$$\begin{aligned} J(W_I, b_I) = &\frac{\lambda_1}{2m} \sum_{i=1}^{m} ||x_i^I - \hat{x}_i^I||^2 \\ &+ \frac{\lambda_2}{2} \sum_{l} (||W_I^{(l)}||_F^2 + ||b_I^{(l)}||_2^2) \end{aligned} \tag{2}$$

where m is the number of samples, λ_1, λ_2 are hyperparameters and $||\cdot||_F$ denotes the Frobenius norm. The first term indicates average error of $\hat{x}^I$. The second term is a weight decay term for decreasing the values of the weights W and preventing overfitting [13].

Stage 2: We train cross edges between two modalities layer by layer. As described in Fig. 2(b), we add correlations between "adjacent" layers from two autoencoders. Formally, we define the cross edges between adjacent layers as:

$$\hat{h}_{I,i}^{(l+1)} = sigmoid(W_{T\to I}^{(l)} h_{T,i}^{(l)} + b_{T\to I}^{(l)}) \tag{3}$$

The cost function to evaluate the difference between $h_{I,i}^{(l+1)}$



(a) Stage 1: basic autoencoders. (b) Stage 2: cross edges. (c) Stage 3: combined weights.

Fig. 2. The details of BDA-CE.

and $\hat{h}_{I,i}^{(l+1)}$ is defined as:

$$J(W_{T\to I}^{(l)}, b_{T\to I}^{(l)}) = \frac{\lambda_3}{2m}\sum_{i=1}^{m}||h_{I,i}^{(l+1)} - \hat{h}_{I,i}^{(l+1)}||^2 + \frac{\lambda_4}{2}(||W_{T\to I}^{(l)}||_F^2 + ||b_{T\to I}^{(l)}||_2^2) \tag{4}$$

Stage 3: We build a neural network with encoder layers and cross edges above, and fine-tune the whole structure. As shown in Fig. 2(c), two pathways are connected and cross edges are added. Each edge is initialized by previous trained parameters. To balance the contribution of ordinary edges and cross edges, we initialize them with a ratio of 0.5. The training target is defined as previously trained autoencoders' middle layer output $h_{I,i}^{(N_h')}$, where $N_h' = \lceil N_h/2 \rceil$. Formally, the recursion formula between adjacent layers is defined as:

$$\hat{h}_{I,i}^{(l+1)} = 0.5 * sigmoid(W_I^{(l)} h_{I,i}^{(l)} + b_I^{(l)}) + 0.5 * sigmoid(W_{T\to I}^{(l)} h_{T,i}^{(l)} + b_{T\to I}^{(l)}) \tag{5}$$

The cost function to evaluate the difference between $h_{I,i}^{(N_h')}$ and $\hat{h}_{I,i}^{(N_h')}$ is defined as:

$$J_I(W_I, b_I, W_{T\to I}, b_{T\to I}) = \frac{\lambda_5}{2m}\sum_{i=1}^{m}||h_{I,i}^{(N_h')} - \hat{h}_{I,i}^{(N_h')}||^2 + \frac{\lambda_6}{2}\sum_{l}(||W_I^{(l)}||_F^2 + ||b_I^{(l)}||_2^2 + ||W_{T\to I}^{(l)}||_F^2 + ||b_{T\to I}^{(l)}||_2^2) \tag{6}$$

The cost function $J_T(W_T, b_T, W_{I\to T}, b_{I\to T})$ is defined similarly. The sum of these two parts of cost is regarded as the final cost function of this stage. After training, the final layers $\langle h_{I,i}^{(N_h')}, h_{T,i}^{(N_h')}\rangle$ are high-level joint representations of the input features, considered as the output of BDA-CE.

Regression Model. To build a mapping from image and text features to aesthetic words in the IAS, we further make the representations $\langle h_{I,i}^{(N_h')}, h_{T,i}^{(N_h')}\rangle$ produced by BDA-CE cast into two-dimensional coordinates $y_i(wc, hs)$. Once we get $y_i(wc_i, hs_i)$ for image v_i, we choose some of the 684 aesthetic words in IAS which has the shortest Euclidean distances with $y_i(wc_i, hs_i)$ as the aesthetic style of v_i. This step can be considered as a regression problem.

3. EXPERIMENTS

3.1. Experimental Setup

Dataset construction. We construct a large benchmark dataset, which employs 120 thousand Flickr images with aesthetic styles. First, using the 684 aesthetic words in the IAS as searched tags, we crawl the Flickr images posted in the last three years. Then, we pick out all the aesthetic words that occur in each image's tags, and calculate the arithmetic mean of their coordinates in the IAS as the aesthetic style ground truth. Finally, we get 119318 (∼120 thousand) social images labeled with aesthetic styles in the IAS.

Feature extraction. 1) *Image features.* Due to the special topic of aesthetics, we extract image features especially from human's aesthetic perception. Referring to [14], we extract 45-dimensional features from images, including figure-ground relationship, color pattern, shape and composition. 2) *Text features.* The text features come from Flickr images' tags, which are independent words. Among the original tags, we remove those occur in the 684 aesthetic words in the IAS, which are used as ground truth. Although the rest tags are not aesthetic style words, they still contain a lot of aesthetic information related to the images, such as "sunshine" and "sea". We adopt Latent Dirichlet Allocation [15] to generate the text features, and set the output dimension to be 30 empirically.

Comparison methods. We compare our model with several baselines from two aspects: 1) connecting different autoencoders to the same regression model DNN (Deep Neural Network) [16], including "None" (no feature learning), "1-DA" (single Deep Autoencoder for both modalities), "2-DA" (one Deep Autoencoder for each modality), and "BDA-CE" (our model), 2) connecting BDA-CE to different regression models, including KNN (K-Nearest Neighbors) [17], SVM (Support Vector Machine) [17], LR (Linear Regression) [16], and DNN.

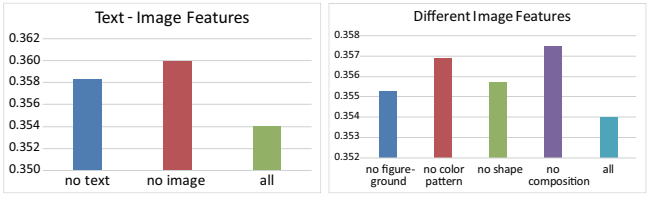
Table 1. Model comparison. (In terms of MSE)

(a) Different autoencoders.

Autoencoder	DNN
None	0.3621
1-DA	0.3590
2-DA	0.3570
BDA-CE	0.3540

(b) Different regression models.

Regression	BDA-CE
KNN	0.3735
SVM	0.3602
LR	0.3598
DNN	0.3540



(a) Text and image features. (b) Different image features.

Fig. 3. Feature contribution analysis.

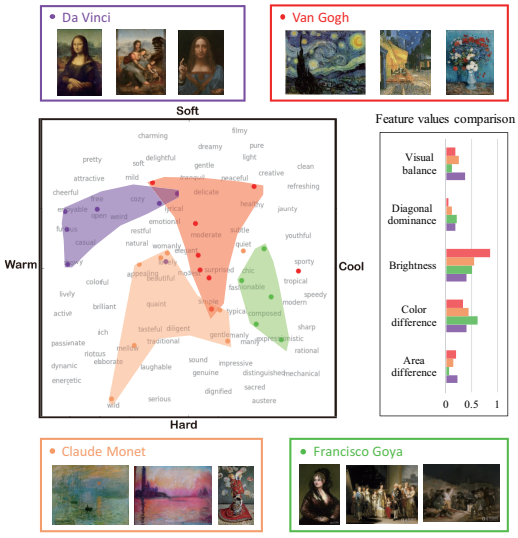


Fig. 4. Aesthetic styles of different artists' works.

Evaluation metrics. We calculate Mean Squared Error (MSE) between predicted coordinates and ground truth, and add up two dimensions' errors as the final metric. All the experiments are performed on five-folder cross-validation.

3.2. Results and analysis

Performance of different autoencoders. Using DNN as the regression model, we compare BDA-CE with other different autoencoder settings. The results are shown in Table 1(a). It is observed that "None" gets the worst performance, proving the usefulness of taking autoencoders as feature learning model. Furthermore, "BDA-CE" performs better than "1-DA" and "2-DA", confirming that our strategy of learning cross edges between modalities to deeply fuse image and text features takes effect.

Performance of different regression models. Using the proposed BDA-CE, we also make several comparisons among different regression models. As shown in Table 1(b), DNN performs better than the other shallow models, probably because deep learning model is more capable of handling the large diversity in our dataset. In the following experiments, we take the best performing DNN as the regression model.

Feature contribution analysis. We compare the contributions of different features separately. First, we discuss the contributions of image and text features. As shown in Fig. 3(a), image features contribute more than text features, which is in accordance with our ordinary feelings. Then we compare the contributions among different types of image features in Fig. 3(b). It shows that color pattern and composition features are more important than figure-ground and shape features, probably due to the characteristics of Flickr images.

Case Study. Employing the images additionally collected from Flickr, we conduct an interesting case study to further show the advantages of our methods. In Fig. 4, we compare the aesthetic styles among different artists' works in the IAS, which are generated from our model, and present the histograms of the five most contributing features. It indicates that Van Gogh's works mainly have a moderate style, probably related to the low color difference between figure and ground. From the perspective of the two dimensions in the IAS, his works present a larger diversity in the hard-soft dimension, but are mostly unified in the warm-cool dimension. Claude Monet's works distribute in separate areas and cover various styles, including natural, typical and mellow. Francisco Goya's works gather in the right part, presenting a unique expressionistic style. Da Vinci's works have lower brightness, tending to be soft and warm and presenting a natural style.

4. CONCLUSION

In this paper, we make an intentional step towards better understanding aesthetic styles of social images. The Image Aesthetic Space we build is a continuous dimensional space which describes the aesthetic styles quantitatively and universally, making it possible to compute aesthetics. The proposed Bimodal Deep Autoencoder with Cross Edges model turns out to be effective for multimodal fusion. We hope that our work can benefit many research and industry fields, such as image retrieval, recognition and classification.

Acknowledgments. This work is supported by the National Key Research and Development Plan (2016YFB1001200), the Innovation Method Fund of China (2016IM010200), the National Natural and Science Foundation of China (61521002), and the Science and Technology Plan of Beijing Municipality under Grant No. Z161100000216147. We would also like to thank Tiangong Institute for Intelligent Computing, Tsinghua University for its support.

5. REFERENCES

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[2] Ming Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi Min Hu, "Global contrast based salient region detection," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 37, no. 3, pp. 569, 2015.

[3] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2285–2294.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[6] Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[7] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller, "Recognizing image style," *Computer Science*, 2013.

[8] Jia Jia, Jie Huang, Guangyao Shen, Tao He, Zhiyuan Liu, Huanbo Luan, and Chao Yan, "Learning to appreciate the aesthetic effects of clothing," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[9] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong, "Towards better understanding the clothing fashion styles: A multimodal deep learning approach," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] Shigenobu Kobayashi, "Art of color combinations," *Kosdansha International*, 1995.

[11] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi, "Wordnet: similarity - measuring the relatedness of concepts," in *National Conference on Artifical Intelligence*, 2004, pp. 1024–1025.

[12] George A. Miller, "Wordnet: a lexical database for english," *Communications of the Acm*, vol. 38, no. 11, pp. 39–41, 1995.

[13] Andrew Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, pp. 1–19, 2011.

[14] Xiaohui Wang, Jia Jia, Jiaming Yin, and Lianhong Cai, "Interpretable aesthetic features for affective image classification," in *IEEE International Conference on Image Processing*, 2014, pp. 3230–3234.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.