Gourmet Photography Dataset for Aesthetic Assessment of Food Images

Kekai Sheng
NLPR, Institute of Automation, Chinese Academy of Sciences & UCAS

Weiming Dong*
NLPR, Institute of Automation, Chinese Academy of Sciences

Haibin Huang
Megvii/Face++ Research US

Chongyang Ma
Snap Inc.

Bao-Gang Hu
NLPR, Institute of Automation, Chinese Academy of Sciences

Humans	☺	☺	☺			☺	
Color + SVM			☺	☺	☺	☺	☺
GIST + SVM	☺	☺			☺	☺	☺
VGG features + SVM	☺	☺	☺			☺	☺
AVA-Networks	☺		☺	☺	☺		☺
GPD-Networks	☺	☺	☺			☺	

Figure 1: Does the food photograph appear aesthetically positive? From top to bottom: human labels, predictions from SVM with handcrafted / VGG features, deep convolutional neural networks based on AVA benchmark [Murray et al. 2012] and the proposed GPD dataset, respectively. The first two images by (Lisa Fotios, rawpixel.com) / Pexles.

ABSTRACT

In this study, we present the Gourmet Photography Dataset (GPD), which is the first large-scale dataset for aesthetic assessment of food photographs. We collect 12,000 food images together with human-annotated labels (i.e., aesthetically positive or negative) to build this dataset. We evaluate the performance of several popular machine learning algorithms for aesthetic assessment of food images to verify the effectiveness and importance of our GPD dataset. Experimental results show that deep convolutional neural networks trained on GPD can achieve comparable performance with human experts in this task, even on unseen food photographs. Our experiments also provide insights to support further study and applications related to visual analysis of food images.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision tasks*;

KEYWORDS

Image aesthetic assessment; Food photography; Convolutional neural networks

ACM Reference Format:
Kekai Sheng, Weiming Dong, Haibin Huang, Chongyang Ma, and Bao-Gang Hu. 2018. Gourmet Photography Dataset for Aesthetic Assessment of Food Images. In *SIGGRAPH Asia 2018 Technical Briefs (SA '18 Technical Briefs), December 4–7, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3283254.3283260

*Corresponding author

SA '18 Technical Briefs , December 4–7, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-6062-3/18/12…$15.00
https://doi.org/10.1145/3283254.3283260

1 INTRODUCTION

Food is one of the most fundamental components in our daily life. The ability to assess the visual aesthetics of food images plays an important role in various tasks, such as food photo selection, recommendation, and post-processing. Although humans can easily gauge the aesthetic perceptions of food photos, performing the same task remains challenging for artificial intelligent systems.

In the past two decades, only a few studies have been explored in related fields, such as image aesthetic assessment [Murray et al. 2012], food categorization [Bossard et al. 2014], photo triage [Chang et al. 2016] and recipe retrieval [Chen and Ngo 2016]. There is very little literature (e.g., [Kakimori et al. 2016]) on visual aesthetics of food images. Beside, currently expert-designed rules cannot cover the complexity of food photos and lack quantitative analysis.

We have two issues to solve at the core of this topic. First, no available dataset can help perform the task and evaluate the solution. Second, prior knowledge is unavailable on how to perform the task sufficiently (e.g., the number of samples required to train a model

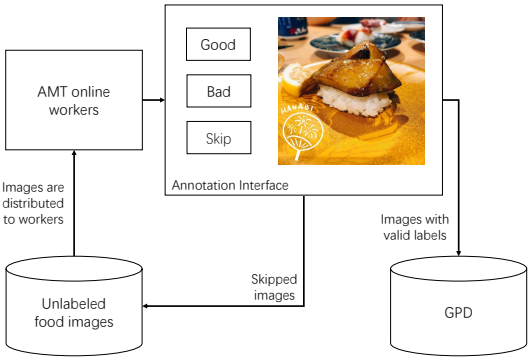

Figure 2: The annotation process of our GPD dataset.

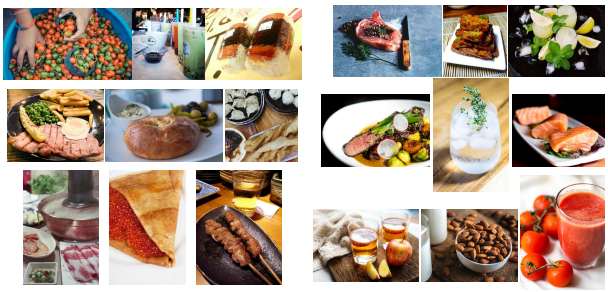

Figure 3: Aesthetically negative (left) and positive (right) food images in GPD. The second/third row of positive photographs by (Oscar Mikols, rawpixel.com) / Pexels.

effectively) and how a model trained on a certain dataset can be generalized to unseen photos. To support research on this topic, we aim at designing an annotation process for a high-quality aesthetic dataset within an acceptable budget.

In this study, we propose the Gourmet Photography Dataset (GPD), which is the first dataset to support the aesthetic visual assessment of food images. We conduct a series of experiments on popular learning mechanisms for visual analysis tasks to verify the annotation quality of the GPD dataset. In addition, we test the generalization abilities of optimized models on unseen instances to demonstrate the effectiveness of GPD. Experimental results demonstrate that GPD provides practical help in tuning models to become predictive of visual patterns that are important to food aesthetics and to realize effective food photo aesthetic assessment. These findings encourage further development in related applications of food photograph aesthetic visual assessment.

2 OUR DATASET

2.1 Data Collection

To learn how to assess the aesthetics of food images, i.e., whether they are aesthetically positive or negative, we aim for high variety in the collected samples: food categories, viewpoint, lighting condition, and layout. Accordingly, we first retrieve related images from social media websites with various food classes and geo-information (e.g., Chinese, French, and Mexican). We also retrieve images from several food categorization datasets (e.g., Food101 [Bossard et al. 2014]) in a class-balanced manner to enrich data complexity. Next, we remove irrelevant samples, such as duplicated images, collages, and images with observable artificial traces. We conduct additional procedures to clean the data, such as removing unnecessary image borders and rotation calibration. Finally, we obtain 16,400 food images.

2.2 Aesthetic Label Annotation

We formally classify food image aesthetic visual assessment as a binary classification problem. For N image-label pairs $\{I_i, \hat{y}_i\}_{i=1}^{N}$, $\hat{y}_i$ is the corresponding aesthetic label for each image I_i. $\hat{y}_i \in \{0, 1\}$ denotes aesthetically negative or positive respectively.

The diagram of the annotation process is shown in Fig. 2. We annotate food images using Amazon's Mechanical Turk (AMT) for aesthetic labels. Workers are asked to provide their judgment on whether the displayed photos look aesthetically pleasing. To ease their anxiety over images with ambiguous aesthetics, workers can opt to *skip* samples if they cannot confidently provide answers. Ensuring the high confidence of answers is crucial for us to limit time consumption and guarantee that labels contain meaningful cues. Images that have been skipped thrice or labeled validly will not be distributed again. Moreover, each worker is allowed to annotate 3,000 images at most. Otherwise, we may risk allowing a few annotators dominate the aesthetic perception of the dataset. A total of 25 workers have participated in annotation procedure, and we obtain 14,968 valid $(I, \hat{y})$ pairs, with 1,432 images skipped.

2.3 Inter-human Agreement

For better quality of GPD, we verify the aesthetic labels and omit wrong/controversial samples as much as possible. Specifically, eight additional expert photographers with good aesthetic perception are invited to observe the annotations. For each image-label pair, they may choose to agree or disagree with the annotated label. Aesthetic labels will be kept if more than four of the experts agree.

After verification, 2, 968 samples are eliminated due to potential controversy. Most of these samples come from several AMT workers who may be unqualified for the task. Eventually, we establish GPD, which contains 12, 000 food images with corresponding aesthetic labels. Fig. 3 shows some of the images in our dataset. In contrast with existing benchmarks on image aesthetic visual assessment (e.g., AVA [Murray et al. 2012]), the proposed GPD focuses on food photographs taken with different types of cameras. Therefore, GPD is arguably a better dataset for supporting research on and applications of food image aesthetic visual assessment.

For simplicity, we randomly divide GPD into two partitions: 9, 600 images (4, 067 negative / 5, 533 positive) for training and the remaining images (1, 016 negative / 1, 384 positive) for testing.

3 MODEL DESCRIPTION

We apply several typical vision learning mechanisms on GPD, as follows. Their performances help us check the quality of labels.

Color + SVM. We compute color histogram features, with 128 bins for RGB color channels. Zero-mean-unit-variance feature normalization is conducted before optimizing SVM.

GIST + SVM. We extract 512-dimensional GIST features with an image size of 256×256. Zero-mean-unit-variance normalization is also performed as a preprocessing step to facilitate training process.

Table 1: Training and test accuracy (%) of different machine learning algorithms and visual features on GPD.

Solution	Training	Testing
Color + SVM	89.06	76.54
GIST + SVM	97.77	77.83
VGG-object + SVM	96.74	88.42
VGG-scene + SVM	93.49	85.71
VGG-food + SVM	96.01	87.71
GPD-AlexNet	77.42	77.25
GPD-VGG	92.22	88.21
GPD-InceptionV2	93.18	89.04
GPD-ResNet	95.78	90.79

VGG features + SVM. We extract 4,096-dimensional representation from the penultimate layer of a 16-layer VGG model [Simonyan and Zisserman 2015]. For comparison, we extract three typical semantics: VGG-object, VGG-scene, and VGG-food, which are trained on ImageNet [Deng et al. 2009], Places365 [Zhou et al. 2017], and Food101 [Bossard et al. 2014], respectively.

GPD-supervised CNNs. We train several popular CNNs, including AlexNet [Krizhevsky et al. 2012], VGG [Simonyan and Zisserman 2015], InceptionV2 [Szegedy et al. 2015], and 16-layer ResNet [He et al. 2016] on GPD. To facilitate optimization and avoid data scarcity issues, all the models are pretrained on ImageNet [Deng et al. 2009]. The loss function is formulated as Equation (1), where $\tilde{y}_i$ is the prediction of a model given I_i, and θ is the trainable parameter:

$$J(\theta) = -\sum_i \log Pr(\tilde{y}_i = \hat{y}_i \mid I_i, \theta). \tag{1}$$

Data pipeline is shared: Photos are rescaled with respect to the shortest edge (259 for AlexNet and 256 for the others), and then patches (227×227 for AlexNet and 224×224 for the others) are randomly cropped. Horizontal mirroring is conducted for data augmentation. We apply different settings of training hyperparameters (e.g., batch-size and learning rate) to maximize each architecture.

4 EXPERIMENTAL RESULTS

4.1 Performance on GPD

In Section 3, we train several vision learning approaches on GPD and their training/validation performance is reported in Table 1. We can obtain two practical information from the results.

- The scale of GPD seems sufficient for supporting the learning processes of the CNNs tested in our experiments. No additional data augmentation or complex training tricks are adopted. The results also indicate the correctness of the proposed GPD.
- GPD-supervised CNNs achieve the best performance. Meanwhile, VGG features generally outperform handcrafted features. These conclusions are in line with the mainstream conclusions.

4.2 Generalization Ability Test

To further validate the effectiveness of GPD, we conduct additional experiments to observe how models optimized using GPD perform in unknown circumstances to test their generalization abilities.

Figure 4: Aesthetic assessment results on unseen food photos. The images on the right side by (Lisa Fotios) / Pexels.

Table 2: Results of several approaches on food photos from WeChat, showing generalization ability quantitatively.

Solution	pos		neg	
	$V(S)$ (%)	$\lvert S\rvert$	$V(S)$ (%)	$\lvert S\rvert$
Best	75.46	296	83.91	529
Worst	16.09	529	24.54	296
Random	37.30	412	62.51	413
Color + SVM	43.24	472	70.43	353
GIST + SVM	48.91	384	72.64	441
VGG-object + SVM	61.43	251	73.12	574
VGG-scene + SVM	56.63	284	72.71	541
VGG-food + SVM	60.03	300	75.55	525
GPD-AlexNet	44.33	579	78.94	246
GPD-VGG	63.30	257	74.33	568
GPD-InceptionV2	62.26	364	75.73	461
GPD-ResNet	71.69	221	75.16	604
GPD-InceptionV2 + GPD-ResNet	66.57	289	78.34	536
Humans (Expert)	72.10	248	81.02	577

4.2.1 Qualitative Evaluation. We collect 2,000 food images from pexels.com, and apply AVA-ResNet (16-layer ResNet trained on AVA [Murray et al. 2012]) and GPD-ResNet to compare their performance on unseen food photos. Fig. 4 presents some results.

GPD-ResNet generally outperforms AVA-ResNet in assessing whether a food image looks good or bad. Although AVA-ResNet can occasionally discriminate good images from bad ones, GPD-ResNet performs food photo triage with higher accuracy. That is, Fig. 4 qualitatively demonstrates that GPD-supervised CNNs exhibit certain generalization abilities, thereby ensuring the necessity and practical values of GPD in food aesthetic visual assessment.

4.2.2 Quantitative Evaluation. To quantitatively measure generalization ability, we collect judgments from 50 qualified candidates on 825 unseen food photos from WeChat. On the basis of their votes, we measure how the aesthetic assessments of optimized models are consistent with human perception using Equation 2:

$$V(S_c) = \frac{1}{|S_c|} \cdot \sum_{I \in S_c} \frac{vote_I^c}{U}, \quad c \in \{pos, neg\}, \tag{2}$$

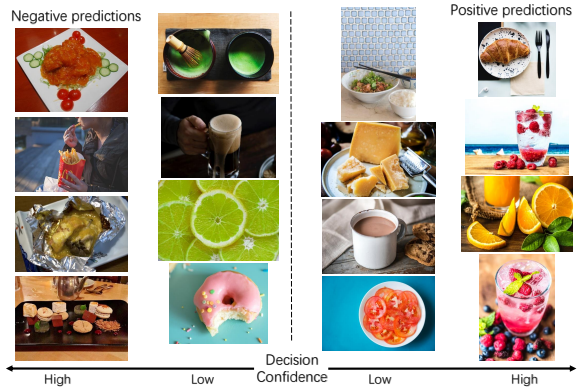

Figure 5: Fine-grained food aesthetics of GPD-ResNet. The images of 2 to 4 columns by (rawpixel.com) / Pexels.

where S_{pos} / S_{neg} denotes which images are positive/negative as predicted by a model, $vote_I^{pos}$ / $vote_I^{neg}$ indicates the number of votes from interviewees who believe I is positive / negative, and U is the amount of candidates for normalization ($U = 50$ in our experiment). The results are listed in Table 2.

We can obtain the following information:

- Aesthetic assessments from GPD-supervised neural networks are consistent with those of human experts. The quantitative results further validate the importance of GPD.
- Negative food aesthetics is easier to model than positive one. Supportive cues arise from the fact that $V(S_{neg})$ is higher than $V(S_{pos})$ across each row. User preferences and data scarcity are the two reasons why modeling positive aesthetics if difficult.

4.2.3 Fine-grained Aesthetic Assessment. Moreover, the proposed GPD-ResNet can generate predictions of three classes that reflect different aesthetic levels: bad, moderate, and good (Fig. 5).

Equation (3) is a simple implementation of this concept.

$$\begin{cases} bad, & p(\tilde{y}_i = 1 \,|\, I_i, \theta) \in [0.0, 0.3], \\ moderate, & p(\tilde{y}_i = 1 \,|\, I_i, \theta) \in (0.3, 0.7), \\ good, & p(\tilde{y}_i = 1 \,|\, I_i, \theta) \in [0.7, 1.0]. \end{cases} \tag{3}$$

Apparently, binary aesthetic annotations with high confidence can provide a driving force to train a fine-grained aesthetic assessment model in a semi-supervised way. These annotations may also verify the correctness of the *skip* operation that we design in annotation procedure. That is, it does not result in the loss of training information to learn effective food aesthetic assessment models.

4.3 Aesthetic Discrimination Visualization

To visualize the important patterns of food aesthetics learned by CNNs, we apply k-means clustering to representations from GPD-ResNet to choose several typical samples, as shown in Fig. 6.

We can leverage these visual patterns to assist in food photography. Systems recommend attractive photographs and offer guidelines to users on how to capture images that are similar to their preferred ones. With such assistances, candidates are encouraged to capture pictures with high quality and increasing variety. In addition, we can flexibly extend the pool of photograph models with images depicting the latest trend.

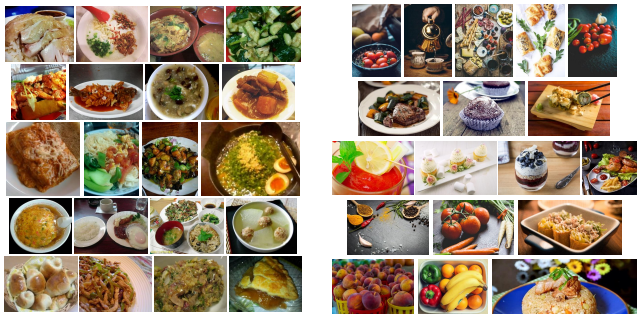

Figure 6: Typical negative (left) and positive (right) photos found by GPD-ResNet, showing a wide variety of patterns.

5 CONCLUSIONS AND FUTURE WORK

In this study, we present GPD, which is the first large-scale dataset for aesthetic assessment of food images. We study the performance of several popular machine learning based algorithms to verify the correctness and effectiveness of our proposed dataset. Experimental results show that the GPD dataset provides valuable help on training models to predict essential visual patterns of food images. We conduct additional generalization tests on unseen food photos to demonstrate that networks trained on GPD can perform comparably with human experts when assessing food photograph aesthetics. In the future we plan to leverage the GPD dataset for various applications related to food photography, including automatic image enhancement, album thumbnail generation, and ranking food photographs for social media needs.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China under nos. 61832016, 61672520 and 61702488, Beijing Natural Science Foundation under No. 4162056, as well as the independent research project of National Laboratory of Pattern Recognition.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *ECCV*. Springer, 446–461.
Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. 2016. Automatic triage for a photo series. *ACM TOG* 35, 4 (2016), 148.
Jingjing Chen and Chong Wah Ngo. 2016. Deep-based Ingredient Recognition for Cooking Recipe Retrieval. In *ACM on Multimedia Conference*. 32–41.
Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE, 770–778.
Takao Kakimori, Makoto Okabe, Keiji Yanai, and Rikio Onai. 2016. A system to help amateurs take pictures of delicious looking food. In *2016 IEEE Second International Conference on Multimedia Big Data*. IEEE, 456–461.
Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.
Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*. IEEE, 2408–2415.
Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR* (2015).
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*. IEEE, 1–9.
Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *T-PAMI* (2017).