Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Image Aesthetics Assessment Based on Multi-stream CNN Architecture and Saliency Features

Hironori Takimoto, Fumiya Omori & Akihiro Kanagawa

To cite this article: Hironori Takimoto, Fumiya Omori & Akihiro Kanagawa (2021) Image Aesthetics Assessment Based on Multi-stream CNN Architecture and Saliency Features, Applied Artificial Intelligence, 35:1, 25-40, DOI: 10.1080/08839514.2020.1839197

To link to this article: https://doi.org/10.1080/08839514.2020.1839197

Published online: 03 Nov 2020.

APPLIED ARTIFICIAL INTELLIGENCE
2021, VOL. 35, NO. 1, 25–40
https://doi.org/10.1080/08839514.2020.1839197

Taylor & Francis
Taylor & Francis Group

Image Aesthetics Assessment Based on Multi-stream CNN Architecture and Saliency Features

Hironori Takimoto[a], Fumiya Omori[b], and Akihiro Kanagawa[a]

[a]Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan; [b]Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama, Japan

ABSTRACT

In this paper, we explore how higher-level perceptual information based on visual attention can be used for aesthetic assessment of images. We assume that visually dominant subjects in a photograph influence stronger aesthetic interest. In other words, visual attention may be a key to predicting photographic aesthetics. Our proposed aesthetic assessment method, which is based on multi-stream and multi-task convolutional neural networks (CNNs), extracts global features and saliency features from an input image. These provide higher-level visual information such as the quality of the photo subject and the subject–background relationship. Results from our experiments support the effectiveness of our approach.

ARTICLE HISTORY
Compiled 23 March 2020

Introduction

The aesthetic quality of an image influences whether a person likes the image intuitively. The purpose of image aesthetics assessment is to predict the perceived quality of an image automatically. Image aesthetics assessment has attracted attention because of its potential use in a wide range of applications, including image retrieval, image cropping, and photo enhancement. However, it is a challenging task owing to its fuzzy definition and highly subjective nature.

Initially, several hand-designed features based methods were proposed to realize image aesthetics assessment as a primary solution (Datta et al. 2006; Dhar, Ordonez, and Berg 2011; Ke, Tang, and Jing 2006; Luo, Wang, and Tang 2011; Marchesotti et al. 2011; Nishiyama et al. 2011). The human judgments given in aesthetic evaluation sets represent human aesthetic experiences and depend on aspects such as colorfulness, contrast, composition, lighting, and subject. Therefore, visual features that contribute to the human perception of aesthetics are manually modeled as low-level features. However, it is difficult to generate precise predictions through hand-designed features based methods, because only a few visual features that contribute to human aesthetic perception are explainable

CONTACT Hironori Takimoto takimoto@c.oka-pu.ac.jp Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Okayama 719-1197, Japan

as explicit knowledge. In addition, it is difficult for hand-designed features based methods to represent high-level semantic information.

The visual recognition paradigm changed rapidly after the appearance of the ImageNet dataset, which demonstrated the power of data-driven feature learning. During the past few years, convolutional neural network (CNN) architectures based on deep learning (LeCun, Bengio, and Hinton 2015) have proven to be the most effective means of facilitating visual recognition. Hence, numerous deep-learning-based approaches for aesthetic assessment of photographs have been proposed (Kong et al. 2016; Lu et al. 2015a, 2015b; Ma, Liu, and Chen 2017; Mai, Jin, and Liu 2016; Omori et al. 2019; Talebi and Milanfar 2018). In previous methods, the problem of aesthetic assessment has typically been cast as a classification or regression problem. The classification problem deals with binary classification, classifying photographs as either high quality or low quality, whereas the regression problem estimates a mean quality score.

The human visual system has a unique ability to selectively focus on the salient and relevant features in a visual scene; this is referred to as visual attention. The core objective of visual attention is to process the least possible amount of visual information when solving complex high-level tasks, e.g., object recognition, helping to ensure that the whole vision process functions effectively. Previous studies have shown that a strong correlation exists between visual attention and visual aesthetics. Coe et al. discovered that aesthetics are a means of drawing attention to an object or person (Coe 1992). In addition, aesthetic objects are interesting, and can thus hold and attract attention (Lind 1980). These studies suggest that visual attention may be a key to aiding aesthetic assessment.

Visual saliency refers to the importance of information obtained from the eyes within the mechanism of visual attention (Treisman and Gelade 1980). Itti et al. (Itti, Koch, and Niebur 1998) proposed a computational model of visual saliency on the basis of Koch and Ullman's early vision model (Koch and Ullman 1985). They demonstrated that a saliency map matches well with the distribution of actual human attention, based on human gaze measurements (Kimura, Yonetani, and Hirayama 2013). In recent years, several CNN-based saliency map estimations have been proposed, and their effectiveness has been demonstrated (Huang et al. 2015; Takimoto et al. 2018).

In this paper, we propose a multi-stream CNN-based image aesthetics assessment method employing saliency features. We assume that a region with high saliency greatly affects the impressions humans gain from photographs, because the gaze is frequently moved to a region with high saliency. In other words, a region with high saliency in a photograph is important in aesthetic assessment. We augment our aesthetic prediction model by adding a saliency feature extraction network based on a multitasking network we previously proposed. The proposed model is a multi-stream network that predicts aesthetics using both global image features pretrained by the ImageNet dataset and saliency features pretrained by a saliency estimation dataset. To more precisely estimate aesthetic quality, the proposed method focuses on biasing the image region that is most important for human impression according to saliency features.

Related Work

The problem of aesthetic assessment has been formulated as a classification problem and as a regression problem. Previous studies have proposed a classification problem that predicts whether a photograph is of high quality or low quality, while others proposed a regression problem that predicts aesthetic quality scores. In the classification task, the top $x\%$ of the mean score is defined as high quality, while the bottom $x\%$ is defined as low quality; in the regression task, the mean score represents aesthetic quality and is directly estimated.

In early research into aesthetic quality assessment of photographs, several methods using various hand-designed features based on professional photography techniques were proposed. Datta et al. proposed automatic classification methods for determining the aesthetic quality of images (Datta et al. 2006). Based on intuition, they designed a number of visual features, including visual cues, wavelet-based textures, and shape convexity. Ke et al. proposed photo quality assessment based on high-level features such as color distribution, simplicity, blur, contrast, brightness, and the spatial distribution of edges (Ke, Tang, and Jing 2006). Dhar et al. proposed different types of human-perceived high-level image attributes related to image configuration, the content of the image, and the natural lighting conditions of the image, to predict image aesthetics and the interestingness of the image (Dhar, Ordonez, and Berg 2011). Luo proposed a content-based photo quality assessment method combined with a set of new subject area detection methods and new global and regional features (Luo, Wang, and Tang 2011). Marchesotti et al. used generic image descriptors to assess aesthetic quality (Marchesotti et al. 2011). They focused on using content-based features, namely GIST (Oliva and Torralba 2001), the Bag-of-Visual-Words (Csurka et al. 2004) and the Fisher Vector (Perronnin and Dance 2007), which encode the distribution of local statistics. However, because of the complexities of human sensibility, it is difficult to sufficiently estimate aesthetic quality using only hand-designed features.

In the past few years, CNNs based on deep learning have achieved state-of-the-art performance on many image recognition tasks. Their deep CNN architectures allow accurate selection of complex, high-level features that are robust to irrelevant input transformations, leading to useful representations that facilitate classification. More importantly, these systems are trained end to end, from raw pixels to ultimate categories, thereby alleviating the need to manually design a suitable feature extractor. CNNs based on deep learning have been used to aesthetics assessment and have shown promising results (Kong et al. 2016; Lu et al. 2015a, 2015b; Ma, Liu, and Chen 2017; Mai, Jin, and Liu 2016). Ma et al. proposed a layout-aware framework in which a saliency map is used to select patches with the highest impact according to the predicted aesthetic score (Ma, Liu, and Chen 2017. Kong et al. proposed a method to aesthetically rank photos by training on the AVA (Esthetic Visual Analysis) dataset with a rank-based loss function (Kong et al. 2016). However, because these methods focus only on mean quality scores, they are insufficient evaluators of aesthetic quality as they do not consider individual differences in human sensitivity. Figure 1 shows an example of two photographs from the AVA dataset, along with their voting distributions (Murray, Marchesotti, and Perronnin 2012). Although the two photographs have an equal mean quality score of 5.5, their content varies significantly. Therefore, it would be incorrect to infer that they are of equal quality.

Talebi et al. proposed an aesthetic quality estimation method that considers the diversity of human sensibility (Talebi and Milanfar 2018). This method, referred to as NIMA (Neural IMage Assessment), directly estimates voting probabilities, which are obtained through subjective quality assessments for each photograph. The mean quality score and standard deviation (S.D.) are indirectly estimated from the obtained voting distribution. Omori et al., which is our research group, proposed a method employing a multi-task CNN to

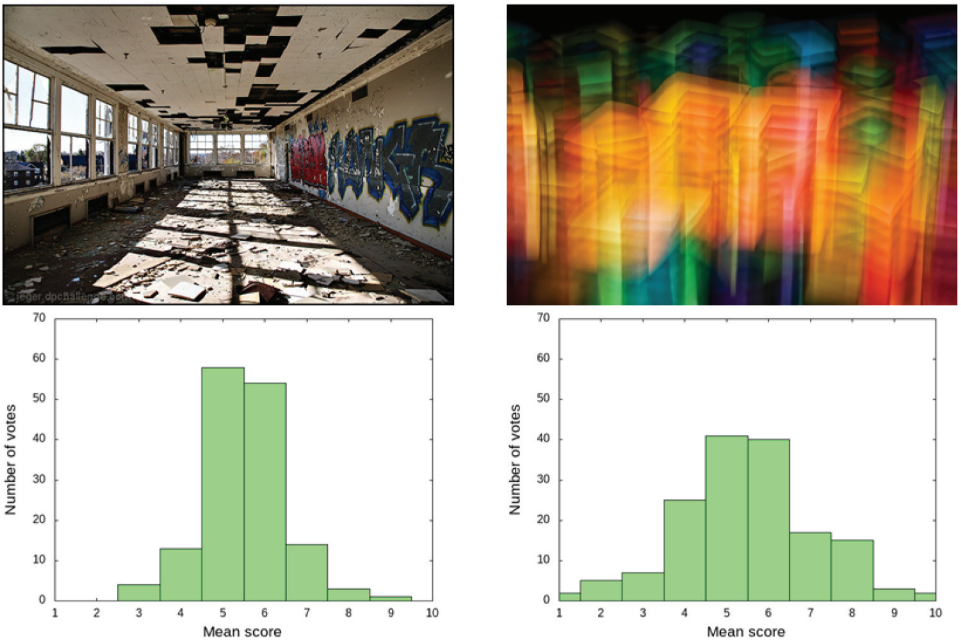


Figure 1. Examples from the AVA dataset: (left) photograph with low variation in its score distribution (S.D.: 0.99), (right) photograph with high variation in its score distribution (S.D.: 1.65).

simultaneously predict the mean score and S.D. of the voting distribution, which better reflects the diversity of the photographs (Omori et al. 2019). The architecture of this method is shown in Figure 2. In this method, Xception architecture is employed as a feature extractor in the main net.

Materials and Methods

Overview

In this paper, we propose an aesthetic quality prediction method based on global visual features and saliency features. Our architecture is shown in Figure 3. The aim of our method employing multi-task CNN architecture is the same as that of our previous method: to more accurately estimate not only the mean score but also the S.D. of the voting distribution indicating photographic quality. Our multi-stream CNN extracts two different features: global visual features and salient features. We employ the Xception model (Chollet 2016) as the backbone of the baseline feature extraction network. Adding the saliency feature extraction network provides higher-level visual

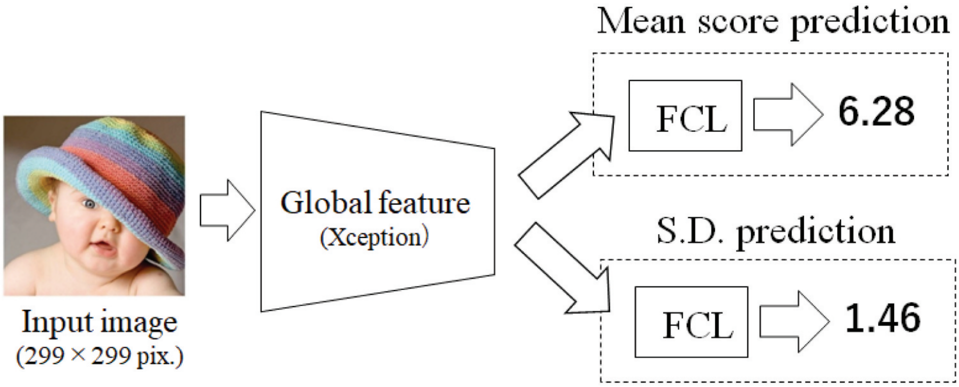


Figure 2. Architecture of aesthetic assessment method based on multi-task learning proposed by Omori et al.

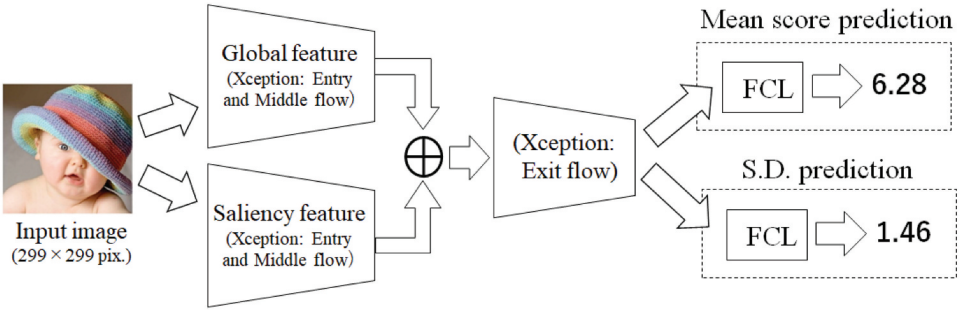


Figure 3. Architecture of the proposed method based on two-stream and multi-task CNNs.

information such as the quality of the photo subject and the subject–background relationship.

Dataset

We employ the AVA dataset, a large-scale database for aesthetic visual analysis (Murray, Marchesotti, and Perronnin 2012). This dataset contains approximately 255,000 photographs that have been aesthetically evaluated. Each photograph has been scored by 200 or more subjects. These scores range from 1 to 10, with 10 indicating the highest quality. The number of votes for each score of each photograph is defined as $V = [v_1, v_2, \ldots, v_{10}]$. v_i represents the number of votes at score i. The mean score is represented as *Score* and S. D. is represented as *SD*. The *Score* and *SD* of each photo are defined as follows:

$$Score = \left(\sum_{i=1}^{10} i * v_i\right) / \left(\sum_{i=1}^{10} v_i\right) \tag{1}$$

$$SD = \sqrt{\left(\sum_{i=1}^{10} (i^2 - Score^2) * v_i\right) / \sum_{i=1}^{10} v_i} \tag{2}$$

However, as a serious issue, the mean scores of the photos in the AVA dataset are biased. Figure 4 shows the distribution of the mean scores in the AVA dataset. In this figure, the number of images for each mean score is shown: the horizontal axis indicates the mean score at 0.1-point intervals and the vertical axis indicates the number of images. The mean scores of images in the AVA dataset follow a normal distribution. In other words, there are few examples of images with very high or very low scores. In a regression problem based on deep learning, the weight of the CNN model is updated based on the error between the estimated value and the true value. As a result, aesthetic prediction models tend to perform poorly when analyzing images with very high or very low scores.

In this study, preprocessing was performed on approximately 255,000 photographs in the AVA dataset, such that the number of distributions per average score was as uniform as possible. First, images such as artificially created illustrations were removed from the dataset. Next, we randomly selected 1,200 images for each 0.1 mean score interval. As data augmentation, a horizontal flip was performed on images in categories with fewer than 1,200 images. Generally, data augmentation operations such as cropping, flipping, scaling, rotating, adding noise, and color transformation are used to preprocess datasets for learning CNN-based models. Enlarging the number of training samples is useful for reducing overfitting and improving generalization. However, image processing associated with data expansion can adversely affect human aesthetic perception. Therefore, we focus only on the horizontal flip as data augmentation. Figure 5 shows the distribution of the

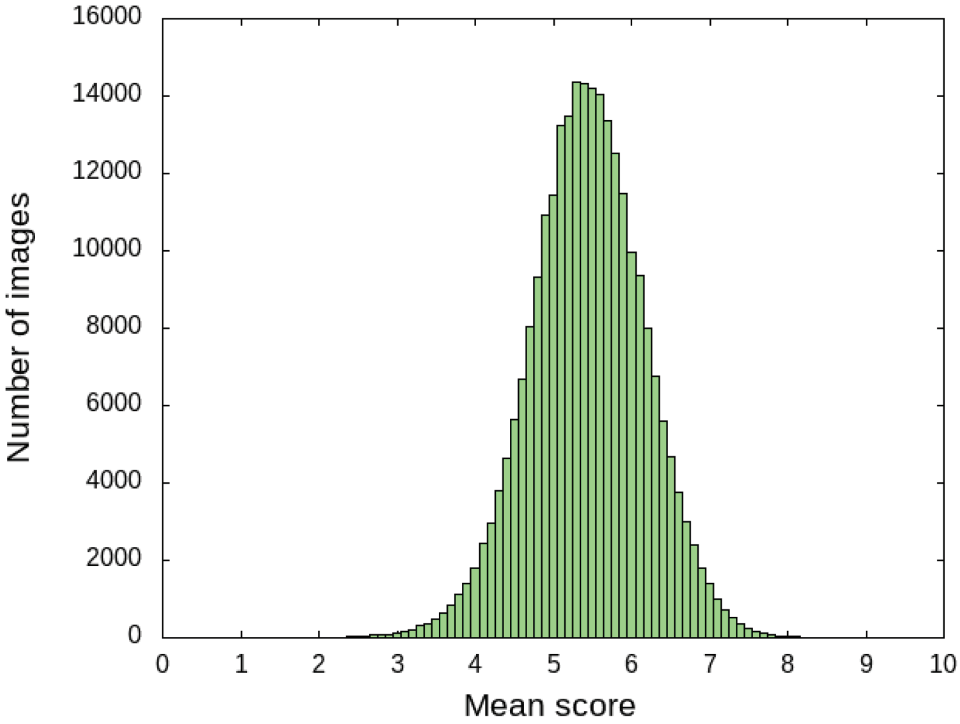


Figure 4. Distribution of mean scores in the original AVA dataset.

mean scores in the uniformized dataset. Note that only the mean score is uninformed, as the main aim of our research is to improve the prediction of the mean score.

Saliency Map Estimation Using Xception

The visual saliency map is a topographically arranged map that represents the visual saliency of a corresponding visual scene. Visual saliency is defined as an estimation of how likely a given region is to attract human visual attention, and there is substantial evidence indicating a correlation between visual attention and saliency maps. It is expected that visual saliency estimation is applied when evaluating prominence in the design of sales promotion tools, public signboards, and so on.

We focus on the visual saliency estimation model described in our previous work (Takimoto et al. 2018). This model precisely estimates a saliency map from images by using the Xception model, which is a state-of-the-art CNN model. The original architecture of the Xception model is shown in Figure 6. Xception architecture is composed of Entry flow, Middle flow, and Exit flow.

As a main contribution in the Xception model, the Inception module is replaced with depthwise separable convolution. The number of parameters and

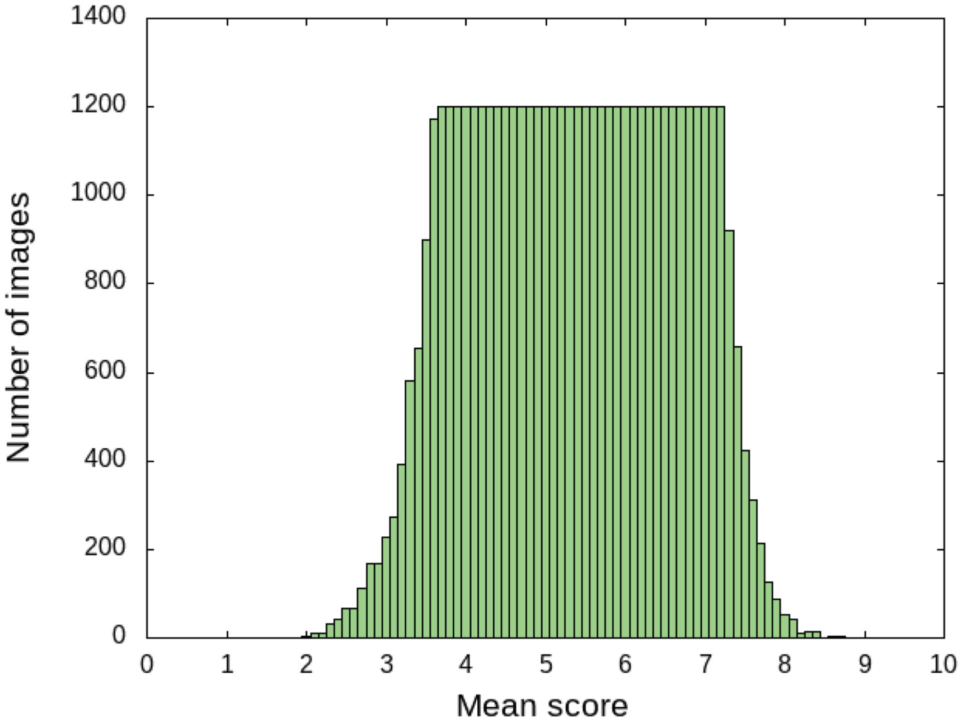


Figure 5. Distribution of mean scores in uniformized AVA dataset.

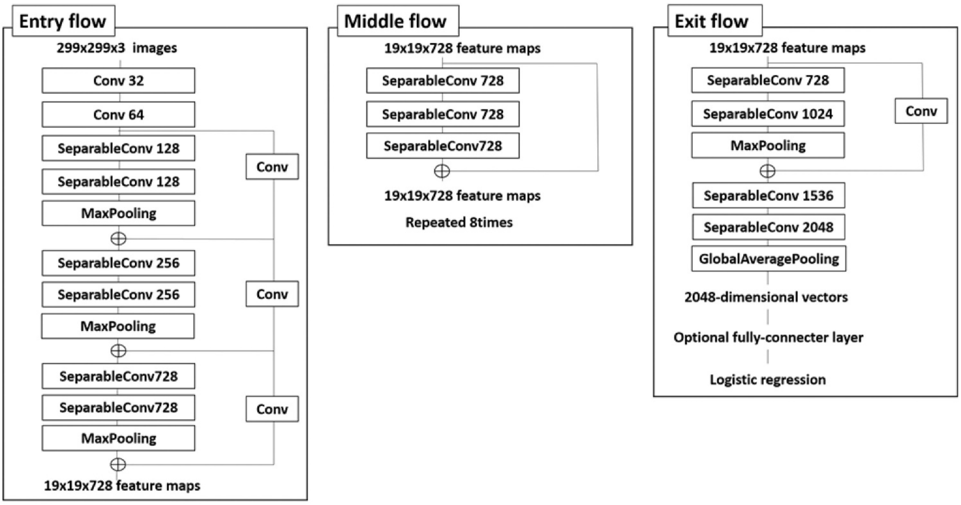


Figure 6. Architecture of the Xception model.

the calculation time are reduced using depthwise separable convolution and pointwise convolution. Depthwise separable convolution is a method for performing spatial convolution of feature maps. Pointwise convolution is a 1 × 1 convolution used in a skip connection such as ResNet, and performs convolution in the channel direction.

The architecture of the saliency estimation model based on Xception is shown in Figure 7 (Takimoto et al. 2018). In our architecture, we used only Entry flow, Middle flow, and Exit flow up through Global Average Pooling (GAP) as the main net for saliency feature extraction. The main net accepts a 299 × 299 RGB color image as input. Let Y_k be a three-dimensional table that contains the responses of the neurons of the CNN at layer k. Y_k has a size of $m_k \times n_k \times d_k$, which depends on each layer. The first two dimensions of the table index the spatial location of the center of the receptive field of the neuron, and the third dimension indexes the templates for which the neuron is tuned. It has been shown that the neural responses at higher layers in the hierarchy encode more meaningful semantic representations than those at lower layers. The last layers of the CNN transform the neural responses to classification scores that have little spatial information. The spatial resolution of the last convolution layer modeling neural responses is 19 × 19. The number of channels at this layer is 728. After feature extraction, as an upsampling network that outputs a saliency map, the transposed convolution layer is used to enlarge the feature map gradually.

Esthetic Quality Assessment Based on Multi-stream Architecture and Saliency Features

In the field of image recognition using CNNs, a model that simultaneously learns and outputs two or more tasks has attracted attention. This model, known as multi-task learning, allows CNN models to share visual knowledge among different attribute categories simultaneously. Each CNN generates attribute-specific feature representations, and then multi-task learning is applied to predict the attributes of the features (Abrar et al. 2015; Sulfayanti et al. 2019).

Meanwhile, multi-stream CNN architecture, in which information from multiple regions is utilized in an additional stream, has recently been adopted

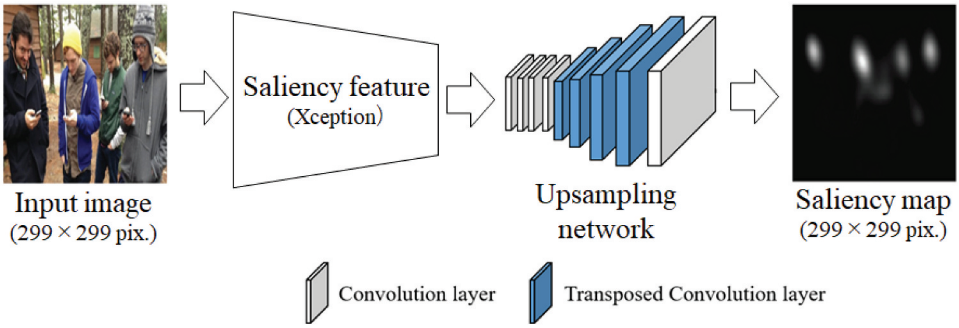


Figure 7. Architecture of the saliency estimation model using Xception.

for use in computer vision tasks. In particular, researchers have proposed action recognition methods based on a two-stream architecture that uses individual frame RGB and optical flow information together with regional features (Chenarlogh and Razzazi 2019; Tu et al. 2018).

We adopted saliency features for quality assessment by using multi-stream architecture. The proposed architecture is shown in Figure 3. First, an image is inputted to two branched streams. Each stream uses the Entry flow and Middle flow of Xception as a feature extractor. One Xception model is initialized using pretrained weights for object recognition. Another Xception model is initialized using pretrained weights for saliency estimation, which is described in section 3.3. Therefore, the visual features branch estimates global visual features contributing to aesthetic quality, while the saliency estimation branch infers visual saliency induced by visual features of the photo.

Then, two different features extracted from the two-stream network are combined and passed to the Exit flow of Xception. The concatenated feature maps is composed of 1,456 channels of size 19 × 19. The portion of the Xception model up through GAP in the Exit flow is used for feature extraction. After feature extraction, this architecture branches into two streams related to a mean score prediction task and S.D. prediction task. Each prediction task comprises a fully connected layer with 256 dimensions, a dropout layer, and an output layer composed of one unit. ReLU (rectified linear unit) is used for the fully connected layer, and a sigmoid function is used for the output layer as an activation function.

The proposed multi-task CNN is trained based on the following loss function.

$$Loss = \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2\} \tag{3}$$

where n is the number of photographs, x_i is the mean quality score of the i th photograph, $\hat{x}_i$ is the estimated quality score of the i th photograph, y_i the S. D. of the i th photograph, and $\hat{y}_i$ is the estimated S.D. of the i th photograph. This loss function for multi-task learning is based on the mean squared error often used as the loss function in regression problems.

Experiments

Experimental Setup

We employed a uniformed AVA dataset, described in section 3.1. A total of 51,376 images were used for training, and 11,150 images were used for testing. All images were resized to 299 × 299 pixels. The initial learning rate was set to 0.0001. All of the models were trained by Adagrad in 100 epochs. The dropout rate was set to 0.5. The batch size was set to 16.

The baseline Xception weights for the global visual feature extractor were initialized by training on ImageNet (Deng et al. 2009), while the baseline

Xception weights for the saliency feature extractor were initialized by training on CAT2000 (Borji and Itti 2015).

As comparison methods, we selected NIMA (Talebi and Milanfar 2018) and a method from our previous work (Omori et al. 2019). By comparing with these methods, the effectiveness of our approach employing saliency features is clearly demonstrated, as the aim of these methods is to predict the mean score and S.D. by using only global visual features. As the main net of NIMA for feature extraction, Inception v2 is employed.

Experiments were performed on a single NVIDIA Titan RTX GPU, using Keras with the Tensorflow backend to implement all models.

Results and Discussion

We evaluated our method with respect to three aesthetic quality prediction tasks: (i) mean score prediction, (ii) score distribution (S.D.) prediction, (iii) aesthetic quality classification. As criteria for prediction tasks (i) and (ii), MAE (mean absolute error), LCC (linear correlation coefficient), and SRCC (Spearman's rank correlation coefficient) were calculated between the ground truth and the predicted mean scores and S.D.'s. The LCC ranges from [−1, 1], with greater values indicating higher correlation. The SRCC evaluates the monotonic relationship between estimated mean scores and ground truth scores. For the classification task (iii), we thresholded the mean scores using threshold t to label the image as having low or high aesthetic quality. We binarized the ground truth and predicted mean scores using a threshold of 5.5, as is standard practice for this dataset.

Table 1 shows the results of three aesthetic quality prediction tasks performed by the compared methods. First, we discuss the results of (i) mean score prediction. All metrics of our proposed method are higher than those of the others. For the MAE results, we performed a pairwise t-test (two-tailed) to

Table 1. Results of three aesthetic quality prediction tasks.

	(i) Mean score			(ii) S.D.			(iii) Classification
	MAE ↓	LCC ↑	SRCC ↑	MAE ↓	LCC ↑	SRCC ↑	Acc.(%) ↑
NIMA	0.689	0.637	0.636	0.198	0.090	0.082	74.3
Omori et al.	0.728	0.565	0.565	0.159	0.150	0.145	71.1
Ours	**0.622**	**0.707**	**0.707**	**0.157**	**0.155**	**0.150**	**78.1**

Table 2. Results of pairwise t-test (two-tailed) for MAE of mean score prediction.

t-value (*p*-value)	NIMA	Omori et al.
Ours	8.40 (< 0.001)	15.75 (< 0.001)
NIMA	–	7.16 (< 0.001)

compare the methods statistically, as shown in Table 2. As a result, we confirmed that the differences between the methods are indeed statistically significant.

Next, we discuss the (ii) S.D. prediction. The metrics indicate that our method performed sufficiently compared with the other methods. For the MAE results, we performed a pairwise t-test (two-tailed) to compare the methods statistically, as shown in Table 3. A statistically significant difference was confirmed between the proposed method and NIMA. However, there was no significant difference between the proposed method and the method of Omori et al.

Finally, we discuss the (iii) aesthetic quality classification. In this table, the results of the classification task are shown as accuracy (%). Although our architecture is modeled for regression, classification accuracy is improved compared with the other methods. We confirmed the effectiveness of using saliency features for aesthetic assessment, as the proposed method achieved the best performance in all tasks.

Figure 8 shows the distribution of the mean absolute error concerning the mean score. The horizontal axis shows the mean score at the 0.1 point interval, and the vertical axis shows the mean absolute error for each score. Although

Table 3. Results of pairwise t-test (two-tailed) for MAE of S. D. prediction.

t-value (*p*-value)	NIMA	Omori et al.
Ours	19.70 ($<$ 0.001)	0.80 ($>$ 0.05)
NIMA	–	18.97 ($<$ 0.001)

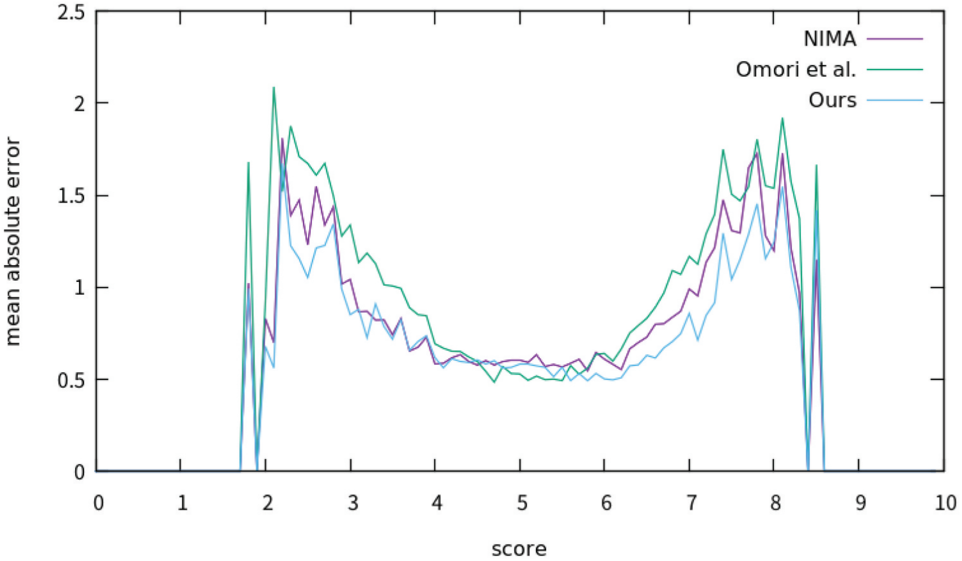
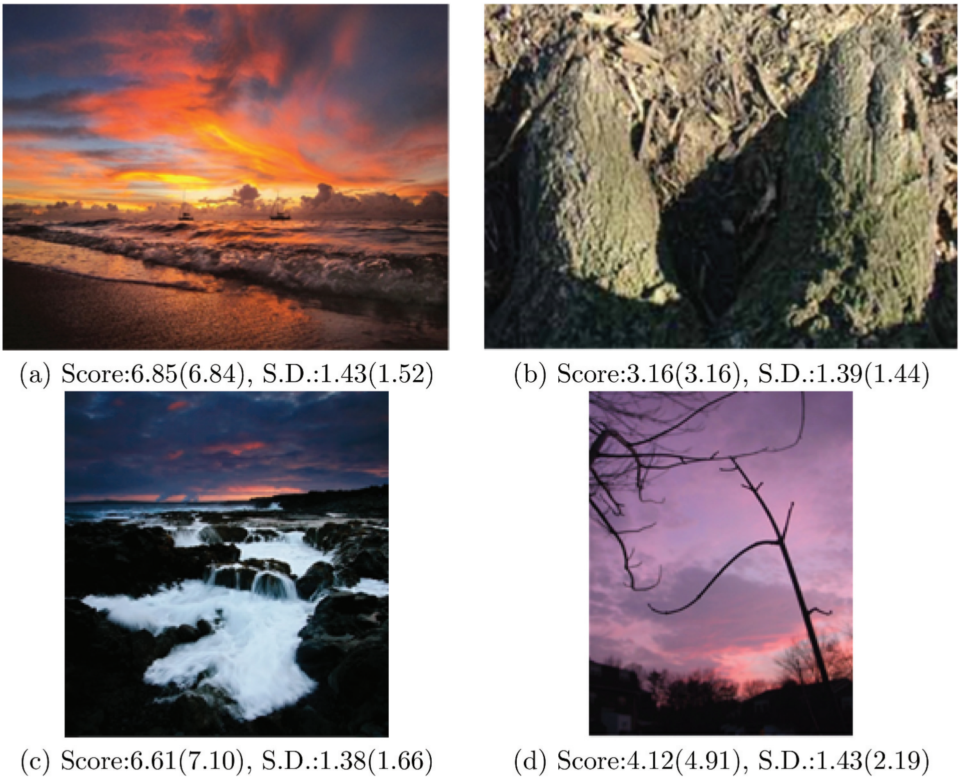


Figure 8. Distribution of the mean absolute error with respect to the mean score.



(a) Score:6.85(6.84), S.D.:1.43(1.52)

(b) Score:3.16(3.16), S.D.:1.39(1.44)

(c) Score:6.61(7.10), S.D.:1.38(1.66)

(d) Score:4.12(4.91), S.D.:1.43(2.19)

Figure 9. Example of estimated result: estimated score and S.D. (ground truth).

dataset preprocessing uniformized the number of images in the dataset, the accuracy for images with high or low mean scores was reduced because there were few images with high or low mean scores. However, the error of our method was smaller than that of the other methods in this situation.

An example of estimated results by our method is shown in Figure 9. In Figure 9(a) and (b), the mean score and S.D. are accurately estimated. Conversely, Figure 9(c) and (d) show examples of insufficient estimation of mean score and S.D., respectively. Images with an average score of 7.0 or greater were rare in the dataset. In addition, for most photographs included in the AVA dataset, S.D. falls within a range of 1.0 to 2.0. Therefore, we assume that it is difficult to accurately estimate Figure 9(d) with a large S.D.

Conclusions

In this paper, we proposed a multi-stream CNN-based image aesthetics assessment method employing saliency features. We augmented our aesthetic prediction model by adding a saliency feature extraction network based on a multitasking network Through comparisons with other aesthetic assessment methods that use only global visual features, we confirmed the effectiveness of using saliency features for aesthetic assessment, as the proposed method achieved the best performance in all tasks.

Funding

This research was supported in part by a Grant-in-Aid for Scientific Research (C) grant from the Japan Society for the Promotion of Science (Grant No. 18K11402).

References

Abrar, H. A., W. Gang, L. Jiwen, and J. Kui. 2015. Multi-task CNN model for attribute prediction. *IEEE Transactions on Multi-Scale* 17 (11):1949–59. doi:10.1109/TMM.2015.2477680.

Borji, A., and L. Itti. 2015. CAT2000: A large scale fixation dataset for boosting saliency research, Proc. of IEEE Int. Conf. on Comp. Vision and Pattern Recog., Workshop on "Future of Datasets. *arXiv Preprint arXiv* 1505.03581.

Chenarlogh, V. A., and F. Razzazi. 2019. "Multi-stream 3D CNN structure for human action recognition trained by limited data". *IET Computer Vision* 13 (4):338–44. doi:10.1049/iet-cvi.2018.5088.

Chollet, F. 2016. Xception: Deep learning with depthwise separable convolutions. Proc. of 2017 IEEE Conf. on Comp. Vision and Pattern Recog., 1800–07.

Coe, K. 1992. Art: The replicable unit - An inquiry into the possible origin of art as a social behavior". *Journal of Social and Evolutionary Systems* 15 (2):217–34. doi:10.1016/1061-7361(92)90005-X.

Csurka, G., C. Dance, L. Fan, J. Willamowski, and C. Bray. 2004. Visual categorization with bags of keypoints. Proc. European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision, 59–74.

Datta, R., D. Joshi, J. Li, and J. Z. Wang. 2006. Studying aesthetics in photographic images using a computational approach. Proc. of the 9th European Conference on Computer Vision, 288–301.

Deng, J., W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition* 248–55.

Dhar, S., V. Ordonez, and T. L. Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. Proc. of 2011 IEEE Conf. on Comp. Vision and Pattern Recog, 1657–64.

Huang, X., C. Shen, X. Boix, and Q. Zhao. 2015. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. *Proceedings of IEEE International Conference on Computer Vision* 262–70.

Itti, L., C. Koch, and E. Niebur. 1998. "A model of saliency-based visual attention for rapid scene analysis". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (11):1254–59. doi:10.1109/34.730558.

Ke, Y., X. Tang, and F. Jing. 2006. The design of high-level features for photo quality assessment. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* 419–26.

Kimura, A., R. Yonetani, and T. Hirayama. 2013. Computational models of human visual attention and their implementations: A survey. *IEICE Transactions on Information and Systems* 96 (3):562––578.

Koch, C., and S. Ullman. 1985. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* 4 (4):219-–227.

Kong, S., X. Shen, Z. Lin, R. Mech, and C. Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. *Proceedings of European Conference on Computer Vision* 662–79.

LeCun, Y., Y. Bengio, and G. E. Hinton. 2015. Deep Learning. *Nature* 521:436–44.

Lind, R. W. 1980. Attention and the aesthetics object. *Journal of Aesthetics and Art Criticism* 39 (2):131–42. doi:10.2307/429807.

Lu, X., Z. Lin, H. Jin, J. Yang, and J. Z. Wang. 2015a. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia* 17 (11):2021–34. doi:10.1109/TMM.2015.2477040.

Lu, X., Z. Lin, X. Shen, R. Mech, and J. Z. Wang. 2015b. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. *Proceedings of IEEE International Conference on Computer Vision* 990-–998.

Luo, W., X. Wang, and X. Tang. 2011. Content-based photo quality assessment. *Proceedings of the IEEE International Conference on Computer Vision* 2206–13.

Ma, S., J. Liu, and C. W. Chen. 2017. A-lamp: Adaptive layout-aware multipatch deep convolutional neural network for photo aesthetic assessment. *Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition* 722–31.

Mai, L., H. Jin, and F. Liu. 2016. Composition-preserving deep photo aesthetics assessment. *Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition* 2016:pp. 497–506.

Marchesotti, L., F. Perronnin, D. Larlus, and G. Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. Proc. of IEEE Int. Conf. on Comp. Vision, 1784–91.

Murray, N., L. Marchesotti, and F. Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. *Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition* 2408–15.

Nishiyama, M., T. Okabe, I. Sato, and Y. Sato. 2011. Aesthetic quality classification of photographs based on color harmony. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* 33–40.

Oliva, A., and A. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42 (3):145-–175. doi:10.1023/A:1011139631724.

Omori, F., H. Takimoto, H. Yamauchi, A. Kanagawa, T. Iwasaki, and M. Ombe. 2019. Aesthetic quality evaluation using convolutional neural network. *Asia-Pacific Journal of Industrial Management* 8 (1):71–77.

Perronnin, F., and C. Dance. 2007. Fisher kernels on visual vocabularies for image categorization. *Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition,* Minneapolis, MN, USA, 1–8. doi:10.1109/CVPR.2007.383266

Sulfayanti, F. S., H. Takimoto, S. Sato, H. Yamauchi, A. Kanagawa, and A. Lawi. 2019. Food constituent estimation for lifestyle disease prevention by multi-task CNN. *Applied Artificial Intelligence* 33 (8):732–46. doi:10.1080/08839514.2019.1602318.

Takimoto, H., S. Katsumata, S. F. Situju, A. Kanagawa, and A. Lawi. 2018. Visual saliency estimation based on multi-task CNN, Proc. of The Fourteenth International Conference on Industrial Management (ICIM2018), Hangzhou, China 639-–643.

Talebi, H., and P. Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27 (8):3998–4011. doi:10.1109/TIP.2018.2831899.

Treisman, A., and G. Gelade. 1980. “A feature-integration theory of attention”. *Cognitive Psychology* 12 (1):97––136. doi:10.1016/0010-0285(80)90005-5.

Tu, Z., W. Xie, Q. Qin, R. Poppe, R. Veltkamp, B. Li, and J. Yuan. 2018. Multi-stream CNN: Learning representations based on human-related regions for action recognition. *Pattern Recognition* 79:32–43. doi:10.1016/j.patcog.2018.01.020.