Aesthetic Quality Assessment of Images Using Text Tags with CLIP

1st Atsushi Sugiura
Department of Robotic Science and Technology,
College of Engineering
Chubu University
Aichi, Japan
tr25009-0170@sti.chubu.ac.jp

2nd Yuji Yamauchi
Department of Robotic Science and Technology,
College of Engineering
Chubu University
Aichi, Japan
yuu@fsc.chubu.ac.jp

***Abstract*—Recently, the aesthetic quality assessment of images has gained significant attention as an important task in computer vision, offering a cost-effective alternative to evaluations performed by expensive experts. Conventional regression-based methods, however, overlook the content of the image, making it challenging to evaluate images with specific intents. To address this issue, this study proposes a novel aesthetic evaluation method that uses both text tags and images. By incorporating the relationship between image content and text tags into the evaluation process, the proposed method enables aesthetic assessment that considers both visual information and semantic elements. Experimental results show that the proposed method yields more accurate and perceptually aligned aesthetic evaluations.**

***Index Terms*—Image Assessment, CLIP, Deep Learning**

I. Introduction

When selecting between photos taken with smartphones or a large number of AI-generated images, users prefer those that are visually appealing and of higher quality. Aesthetic quality assessment has become a common method for evaluating visual beauty and appeal.

A method for predicting aesthetic preference scores assigned by humans via regression was proposed in [1]. This approach directly predicts the score using only the visual features of the image, but it struggles to capture complex visual elements and the underlying content of the image.

Therefore, in this study, we propose an aesthetic quality assessment model that uses text tags representing image content. The proposed method adopts CLIP [2] as the base model and integrates both text tag information and visual features during training. Furthermore, we train the model to evaluate image beauty based on ranking, using both human-perceived image beauty and the semantic similarity between the image and its content. The model enables evaluations that reflect both visual beauty and content relevance better aligning with user needs. Our contributions are: (1) Demonstrating the effectiveness of integrating text tags, and (2) showing improved alignment with human perception.

II. Proposed Method

A. Training Approach

The proposed method uses images and text tags that represent their content, allowing the evaluation of images to consider these tags. Furthermore, by introducing ranking learning based on the consistency between the image and its corresponding tag, the model can assess image superiority based on predetermined criteria for each text tag.

Figure 1 illustrates the training flow of the proposed method. We adopt the CLIP [2] model as the base. CLIP is a multimodal model that learns paired relationships between language and images through contrastive learning.

First, CLIP takes two pairs of images and their corresponding text tags as input and generates 512-dimensional embedding vectors for each image and each text through their respective encoders. CLIP is a pre-trained model. Next, these vectors are passed to a cross-attention module, where image and text information are jointly learned, and a 512-dimensional feature vector is extracted for each image-text pair. Finally, a multilayer perceptron (MLP) processes the embeddings to output a score that represents the aesthetic quality of the image.

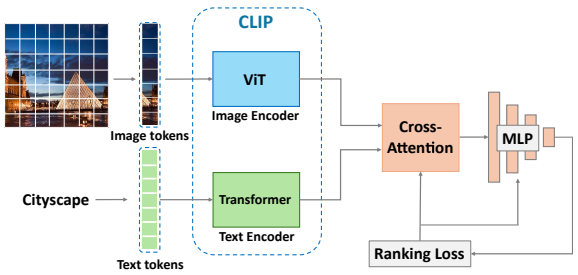

Fig. 1. Training flow of the proposed method.

B. Definition of ranking information

To maintain accuracy in aesthetic score prediction while improving ranking consistency between images, we introduce ranking learning. For images sharing the same text tag, we provide ranking information using both aesthetic scores and CLIP similarity as criteria. This allows the model to rank images higher if they are aesthetically superior and more semantically aligned with the text tag.

Aesthetic scores are the average of evaluations from multiple people and numerically represent the image's beauty.

CLIP similarity quantifies the semantic relationship between the image and its tag. Specifically, images are initially ranked by aesthetic score. When multiple images have similar scores, those with higher CLIP similarity are ranked higher.

C. Loss Function

The proposed loss function $L(\theta)$ consists of two components: a ranking-based loss and a regression loss. It is calculated based on the predicted score difference between a higher-ranked image (T, x_i) and a lower-ranked image (T, x_j), and is used to update the weights of the cross-attention and MLP. The loss function is shown in Equation (1):

$$\begin{aligned} L(\theta) &= -\mathbb{E}_{(T,x_i,x_j)\sim D}\left[\log\left(\sigma\left(f_\theta(T,x_i) - f_\theta(T,x_j)\right)\right)\right] \\ &+ \lambda \cdot \frac{1}{D}\sum_{i=1}^{D}\left(\left(f_\theta(T,x_i) - s_i\right)^2\right) \end{aligned} \tag{1}$$

Here, T is the text tag, x is the image, f_θ is the score for the text tag and image, σ is the sigmoid function, D is the number of data points, s is the ground truth score for the image x, and λ is a weight used to balance the components.

The first term in equation (1) represents the ranking-based loss, while the second term represents the regression-based loss.

In the proposed method, images with the same text tag are compared, and based on the ranking information, optimization is performed so that the predicted score for image x_i is higher than that for image x_j.

By directly predicting the score and incorporating the ranking loss term, it becomes possible to learn the relative quality of images based on the given metric.

III. Experiments

A. Experimental Setting

To verify the effectiveness of the proposed method, we conduct experiments using the Aesthetic Visual Analysis (AVA) dataset [3]. The AVA dataset includes images annotated with average aesthetic scores from multiple individuals, along with text tags representing image content. We use 135,000 image-tag pairs for training and 1,350 pairs for evaluation. The average aesthetic scores in AVA serve as ground truth, and we compare the aesthetic assessment accuracy of the proposed method with that of a conventional method [1].

B. Experimental Results

Table I shows the inference accuracy of aesthetic evaluation. As shown, the proposed method achieves similar error rates comparable to conventional methods but outperforms them on all other metrics. This suggests that the proposed method improves the performance compared to the poor regression results observed in earlier methods.

Examining the KL Divergence and EMD between the predicted and actual score distributions, the proposed method better approximates the distribution of human-assigned scores. Figure 2 shows overlapping distribution plots for the predicted and ground-truth scores for 1,350 evaluation images. The results indicate that conventional methods were influenced by data skew, whereas the proposed method maintains consistent evaluation standards.

TABLE I
AESTHETIC EVALUATION RESULTS OF EACH METHOD:THE ERROR AND ERROR VARIANCE ARE CALCULATED BASED ON THE PREDICTED VS. GROUND-TRUTH SCORES. LINEAR CORRELATION COEFFICIENT (LCC) AND SPEARMAN RANK CORRELATION COEFFICIENT (SRCC) ARE COMPUTED BETWEEN THE PREDICTED AND ACTUAL AVERAGE SCORES.

Method	Error	LCC (mean)	SRCC (mean)	KL Divergence	EMD
Previous	0.478(±0.36)	0.596	0.556	0.295	0.106
CLIP+MLP	0.442(±0.136)	0.650	0.581	0.185	0.116
Ours	0.482(±0.147)	0.684	0.620	0.138	0.094

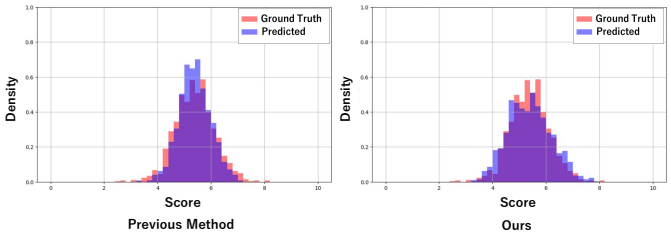

Fig. 2. Distribution of scores for each method.

Figure 3 shows example aesthetic evaluation scores for a single image. The conventional method, which only considers visual features, diverges from human evaluation. In contrast, the proposed method, which incorporates text information (e.g., "Birds"), produces a score much closer to human judgment.

TextTag	Birds
Previous	5.36
Ours	6.62
Human Evaluation	6.84

Fig. 3. Examples of images, tags, and aesthetic scores.

IV. Conclusion

In this study, we proposed an image aesthetic quality assessment model that incorporates text tags. In future work, we aim to develop models that more closely mirror human aesthetic evaluations.

References

[1] Talebi, *et al*. "NIMA: Neural image assessment." IEEE transactions on image processing, 2018.

[2] Radford, *et al*. "Learning transferable visual models from natural language supervision." ICML, 2021.

[3] Murray, *et al*. and Florent Perronnin. "AVA: A large-scale database for aesthetic visual analysis." CVPR, 2012.