Understanding Scenery Quality: A Visual Attention Measure and Its Computational Model

Yuen Peng Loh†, Song Tong, Xuefeng Liang, Takatsune Kumada, Chee Seng Chan†
IST, Graduate School of Informatics, Kyoto University, 606-8501 Kyoto, Japan
† Centre of Image and Signal Processing, University of Malaya, Kuala Lumpur, 50603 Malaysia
lohyuenpeng@siswa.um.edu.my, tong.song.53w@st.kyoto-u.ac.jp
xliang@i.kyoto-u.ac.jp, t.kumada@i.kyoto-u.ac.jp, cs.chan@um.edu.my

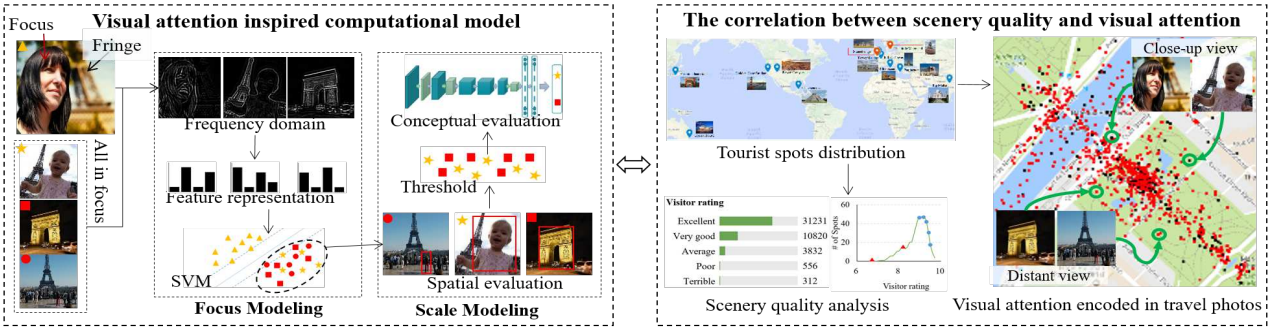


Figure 1: Measure the scenery quality of a tourist spot by analyzing tourists' visual attention.

Abstract

Travel photos record tourists' experiences and attentions when visiting a place. We question if they embed any untapped indices, subconsciously created by the tourists, for measuring the scenery quality? By analyzing thousands of such photos and inspired by the psychological theory of "broaden-and-build", our study reveals a strong inclination of taking panoramic photos at high rating outdoor tourist spots. Thus, this preference can be a supplementary measure of indexing the scenery quality. However, the task of recognizing panoramic photos is nontrivial. In this paper, we propose a visual attention inspired computational model to address this issue, which mimics human perceptual and cognitive mechanisms by a focus model and a scale model. The experiments on a newly created dataset demonstrate a remarkable performance of our proposal, along with its effectiveness in measuring scenery quality also verified by 10 high rating outdoor spots and 2 lower rating ones from across the world.

1. Introduction

Recent years have witnessed a revolution in internet Big-Data for tourism economics. It has enabled us to accumulate massive volumes of travel data from social networking services (SNSs) and travel guides, and generates valuable knowledge for tour recommendations [4, 18, 24, 35]. To provide a precise recommendation to the end users, one important task of these online services is to evaluate the "scenery quality" of each tourist spot. The common methodologies employ the "rank-by-count" or "rank-by-frequency" using check-ins, tweets, or GPS traces shared through SNSs [18, 20, 22, 23, 44]. Nevertheless, these methods are arguably over-optimistic because people often open positive opinions in their tweets and check-ins but are apt to hide negativities, and GPS traces are somehow personalized. We question if there exist any untapped indices that unveil tourists' subconscious preferences in spite of all disguises?

After browsing thousands of travel photos collected from both high and lower rating tourist spots, we find a considerable inclination of tourists to take more distant/panoramic view photos at high rating spots, but this preference appears to be moderate or even going in the reverse direction at lower rating spots, see Fig. 6. This observation follows an inspiration from the well-established psychological theory of "broaden-and-build" [16, 33, 34], where positive emotions instigate a preference to a global visual attention but negative emotions narrow down the attention. Thus, the reasonable explanation of the observation is that people would have positive emotions and tend to broaden their attention when experiencing high quality spots. Transferred into action, they could be shooting more distant/panoramic view photos to record their broader attention for memory [1, 9].

On the contrary, tourists would focus more on small elements at the lower rating spots to capture close-up view photos. In this phenomenon, we consider that a tourist's emotion correlates with the user rating at trip recommendation sites (*e.g*. TripAdvisor) which indicates the scenery quality of the spot. Thus, evaluating the scenery quality of an outdoor spot can be partially treated as a problem of measuring tourists' attentions at that place, and further simplified as an image classification task of estimating the proportion of those distant/panoramic view photos.

However, images shared by tourists through SNSs have an extremely wide variety of contexts that greatly challenge this task. In this work, we follow the basic idea of human visual system (HVS) and propose a framework consisting a *focus model* and a *scale model*. The *focus model* is based on the finding that a large number of professionally shot close-up view photos adhere to the focus lens model of HVS [41] where it focuses on the center object (focus) while the surrounding background is blurred (fringe), as shown in Fig. 2(a). To model it, we transform images into a set of different frequency domains, and applied three well accepted feature descriptors and two codebook approaches, respectively. Afterwards, a support vector machine classifier is built to find the best domain and feature to represent this focus model. However, many close-up view photos shot by low-cost cameras (*e.g*. smart phones) do not follow the focus model where entire scene appears sharp, as shown in Fig. 2(b). Therefore, the *scale model* is derived from observers' ability to differentiate the views by measuring the size of objects, namely the *spatial size* (the object size measured in the photo indicated by the boxes in Fig. 2(b) and 2(c) is bigger than the one in Fig. 2(d)) and *conceptual size* (the realistic proportion of the object; a person in Fig. 2(b) is a small object and a building in Fig. 2(c) is a big object). We measure the spatial size by five variations of object bounding box proposal methods, whereas, the conceptual size is measured by object recognition (a fine-tuned CNN). Then, the best combination of box proposal method and recognition CNN is selected. Finally, both the focus and the scale models are assembled to form one framework.

To validate the effectiveness of our computational model, we collected 5050 images from Flickr and ImageNet including 2452 distant/panoramic view and 2598 close-up/local view photos. Experiment results showed that our proposed framework achieves a noticeable improvement from 84% [39] to 93.17% in overall accuracy (Table 3). Furthermore, we statistically analyze photos taken from 10 high rating spots and 2 lower rating ones on TripAdvisor. The result reveals a strong correlation between the viewpoint ratios with the tourists' ratings of these spots. Particularly, the ratio approximates 8:2 or 7:3[1] in favour of panoramic

[1]Surprisingly, many online tour recommendation articles [15, 26, 37] subconsciously use this ratio to illustrate nice locations as well.

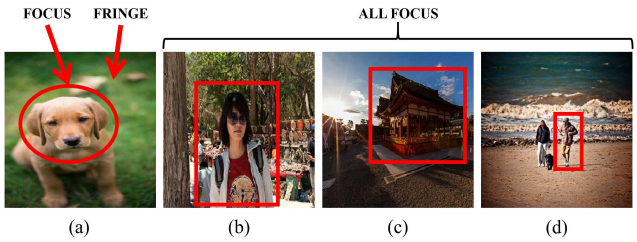


Figure 2: Image samples. (a) Close-up view with focus and fringe; (b) All focused close-up view with large spatial but small conceptual size object; (c) All focused distant view with large spatial and large conceptual size object; (d) All focused distant view with small spatial and small conceptual size object.

views at high rating spots. We boldly claim this inclination as a supplementary measure of spot scenery quality established on top of the subconscious behavior of tourists due to the consistency with the psychological studies on visual attention.

The summary of this work is illustrated in Fig. 1, and the contributions are threefold.

- We investigate various computational algorithms to build the HVS inspired focus model and scale model for a reliable framework to perform the distant/panoramic view and close-up/local view classification task.
- We perform a statistical analysis using our proposal on 24K photos taken from a dozen outdoor tourist spots, and found a strong correlation between the viewpoints of travel photos with the quality ratings on *TripAdvisor*.
- We proposed a scenery quality measure based on the aforementioned analysis with the support from literatures about the psychology of human visual attention.

2. Related Works

The preference of panoramic views at high rating spots could be explained by a list of psychological studies on visual attention. It has been known that the interpreted content of a scene can be viewed as two levels in the human perceptual process; that is to say *global* processing and *local* processing. Navon [27] claimed that human attends to process global structure of a scene or fine-grained elements according to varied tasks, however, a global precedence could presumably hold when both global and local levels have the same visibility. Other researches further illustrated that emotions could interact with visual attention and affect perceptual process [2, 8, 17]. Specifically, Fredrickson *et al*. proposed a "broaden-and-build" theory [16]. It shows that positive emotions broaden (globalize) the scope of attention of the observer and result in processing of a global picture, while negative emotions correlate with a narrowed

(localized) attentional focus and induce the processing of local units of the presented stimuli. Conversely, Niedenthal *et al*. [28] demonstrated that distributed attention leads to positive emotions and focused attention leads to negative emotions. These conclusions show that emotional and perceptual processes interact reciprocally [11, 14, 33, 34].

To our best knowledge in computer science, very few works are aware of this psychological phenomenon and its potential applications. Only a few researches address on the problem of camera viewpoint. Zhuang *et al*. [45] employed the edge distribution based on an assumption that panoramic images have gentle contrast throughout the whole image, but other images do not have. However, the irregular spatial information would degrade the accuracy. Torralba *et al*. [40] investigated the relationship between the image structure and spectral signatures in frequency domain. They proposed an image feature based on discrete Fourier transform (DFT) for classification. Unfortunately, DFT can not fully represent the focus attribute of a close-up view [30].

Our previous work [39] explored the *focus cue* and *scale cue* using the Discrete Wavelet Transform and the Edge Box, respectively, and achieved a reasonable performance on a relatively small dataset. Instead, this work investigates a large variety of methods, and demonstrates a noticeable improvement from 84% [39] to 93.17% in overall accuracy on newly created dataset. Moreover, we link the scenery quality with the preference of photo-taking, and suggest a supplementary index.

3. The Computational Model

The task of distant/panoramic and close-up/local view classification is not as simple as it seems due to context variability. Therefore, we approach the problem based on human visual perception by investigating a model that closely resembles the HVS. In psychology, the visual attention operation is described by the focus lens model [41] whereby it defines the focus, fringe, margin, and size changing. As shown in Fig. 2(a) the *focus* is the area central to the visual field where high-resolution (*i.e*. sharp) information are extracted, whereas the *fringe* is the area surrounding the focus where low-resolution (*i.e*. blur) information are derived. The fringe extends to the margin where the vision field ends, while the size changing describes the trade-off in the processing efficiency when the region of focus varies. The interest point of establishing a computational model that replicates this psychological model, particularly the properties of focus and fringe, is the ability to understand the viewpoint of images for subsequent scenic evaluations.

3.1. Focus Modeling

The focus and fringe properties are found in close-up view photos, particularly the ones taken by professional photographers, as a result of the shallow depth-of-field (DoF) settings in camera lens. Hence, these image properties can be exploited in the frequency domain where the high pass subband contains the sharp details of the image, while the blur and smooth textures are found in the low pass subband. Figure 3 shows examples of images transformed to frequency domain, where it is apparent that the low frequency consists mostly of the image background which is noticeably blurred compared to the grayscale image. On the other hand, the high frequency decomposition shows sharp edge details of the image. The discriminating factor is that the close-up view has high frequency details that are concentrated towards the center of the image, whereas the details of distant view images are scattered.

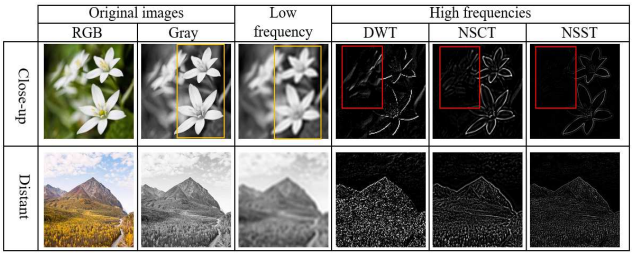


Figure 3: Example of close-up and distant view images transformed into frequency domains using DWT, NSCT, and NSST. (Best viewed in color.)

3.1.1 Frequency Domain and Feature Representation

There are various frequency transformation methods available such as the Discrete Wavelet Transform (DWT) [30], a simple, flexible and fast multi-resolution decomposition approach. Differently, contourlet [7, 10] and shearlet [12, 19] are more sophisticated methods that can better represent sharp and blur information. Contourlets and shearlets are anistropic transformations which give directional sensitivity that wavelet lacks. The nonsubsampled variant of them, the Nonsubsampled Contourlet Transform (NSCT) [7] and the Nonsubsampled Shearlet Transform (NSST) [12], further introduces shift invariance by eliminating the downsampling and up-sampling.

In our investigation, the high frequency decompositions of DWT, NSCT, and NSST as shown in Fig. 3 has clear differences that can influence the classification process. It can be seen that there are background details "leaked" into the decomposition of DWT bounded by the red box. But it is not apparent for the decompositions by NSCT and NSST, hence are better representations of focus information.

With the established frequency domain to emphasize the sharpness information of the images, the next step is to extract features to represent the different views. The multi-windows based histogram of frequency energy (MWHFE) [39] approach quantizes the pixel-wise energies in the high

frequency decompositions. Designed for the high frequency domain, it captures the spatial distribution of the sharp details produced by the DWT, NSCT, and NSST, as features.

Additionally, we implement other approaches as comparison to investigate the most effective approach for modeling the focus cue. Two notable image feature extraction methods were tested, namely the Local Binary Pattern (LBP) [29] and Speeded-up Robust Features (SURF) [3] that captures texture and edge features respectively. The texture difference of sharp and blur greatly differ from one another, hence the LBP is suitable for the classification task. Conversely, SURF would be able to capture the apparent variations of the edges in center focused close-up and all focus distant view images. As per convention, the implementation of local features such as SURF in classification includes codebook feature quantization, therefore, we further investigate the effectiveness of different codebook approaches for our task. The methods we engaged are the Bag of Visual Words (BoVW) [6] and Fisher Vector (FV) [31, 32]. A Support Vector Machine (SVM) is chosen as the focus feature classifier because of its robustness on binary classification problems.

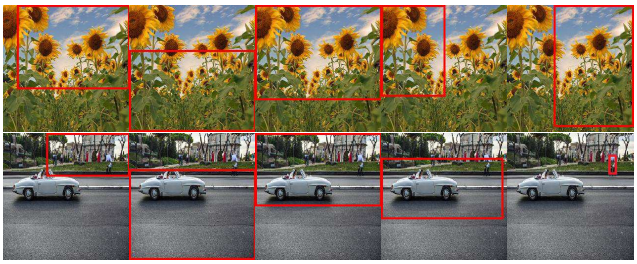
3.2. Scale Modeling

We found that a large portion of close-up view photos, which are misclassified as distance view, do not have the focus and fringe characteristics. Specifically, Fig. 2(b) appears all focused in whole image, thus, confuses the focus model for classification. One of the reason for this is due to the advancement of smart phone technology enabling many tourists to use the inbuilt compact cameras, which have large depth-of-field, to take photos while traveling.

Nevertheless, we realize that the distance of objects from the viewpoint results in different sizes in photos. This is a crucial characteristic where we can distinguish close-up and distant view by evaluating the object size in the photo (spatial size) and the object's realistic scale (conceptual size). In close-up views, conceptually small objects (*i.e.* people, dogs) has large spatial size as shown in Fig. 2(b), while, these same objects are spatially small in distant views, see Fig. 2(d). Therefore, measuring the spatial size can be an indicator of the view type. However, conceptually large objects (*i.e.* buildings, mountains) can have big spatial size, as seen in Fig. 2(c), similar to conceptually small objects in Fig. 2(b). Such confusion can be eradicated by evaluating their conceptual sizes.

3.2.1 Spatial and Conceptual Size Evaluation

Object bounding box proposers can be exploited to evaluate an object's spatial size without recognizing the exact object within the image. The spatial size can be approximated efficiently by checking the size of the proposed object bounding box determined by the corresponding objectness score. If a high scoring box is found to be smaller than a predetermined threshold, the image can be classified as a distant view, such as Fig. 2(d).



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

Figure 4: Example of bounding box proposed using (a) Edge Boxes, (b) Adobe Boxes, (c) AdobeBING, (d) RPN (ZF) and (e) RPN (VGG16). (Top: Close-up view image; Bottom: Distant view image.)

Object proposal algorithms have gained great interest as a means to speed up object detection tasks into real time systems. Consequently, many object proposal methods have been introduced that is constructive for the spatial size evaluation task. We look into the Edge Boxes [46], Adobe Boxes [13], and Region Proposal Network (RPN) [36], each approach based on different sets of assumptions.

The Edge Boxes proposes object bounding boxes based on the grouping of edges, and uses the edge content of the bounding box to compute objectness (likelihood it is an object) score. Whereas, the Adobe Boxes uses the collection of superpixel with high color contrast from the background as the representation of object parts, named adobes, to localize objects and the spatial compactness of these adobes are used to calculate the objectness score. Additionally, the Adobe Boxes can be used as a refinement for other proposal algorithms, where it is recommended to be used with the Binarized Normed Gradients (BING) [5]. While BING is an object proposal algorithm by itself, the given objectness scores are biased towards full image, hence, interferes with the spatial size evaluation. Therefore, the suggested Adobe refined BING (AdobeBING) is included in the investigation instead of BING itself. As for RPN, it is a deep learning approach using fully convolutional network (FCN), designed for speeding up convolutional neural networks (CNN) by sharing the convolution parameters of a specified object detection network. It is an end-to-end learning based approach that does not require feature designs such as edge for Edge Boxes, and superpixels for Adobe Boxes. As the RPN is detector specific, two variations were tested, the RPN based on the Zeiler and Fergus' CNN model (ZF) [43], and another based on Simonyan and Zisserman's CNN model (VGG16) [38]. Although designed to be paired with CNN object detectors, the RPNs are implemented as a

standalone object proposer because the identity for the objects are not necessary for our evaluation, and also to set up a fair comparison with the other proposal methods.

Figure 4 shows examples of object bounding boxes proposed by the aforementioned methods. Evidently, these methods give vastly different object focus and box shapes for the same images. Hence, it is necessary to investigate the best option for an adequate view classifier.

Please note that the object proposal size evaluation can only filter out distant views where conceptually small objects (*e.g.* people, animals) having small spatial size, such as Fig. 2(d). However, the cases in Fig. 2(b) - 2(c) must be handled by another classifier to evaluate the conceptual size of the object in the proposed bounding boxes, where large is categorized as distant view, while small as close-up view. To this end, a generic CNN object classifier is chosen and fine-tuned into a binary classifier for the final stage of measuring object conceptual size.

4. Experiments

The experiments of the computational model were conducted using a dataset of close-up/local view and distant/panoramic view images we have collected from the Flickr and the ImageNet database. The dataset contains 2598 close-up view images and 2452 distant view images. This dataset is then divided into subsets for the experiments. The first subset, referred as Set1 henceforth, consists of 1522 close-up images that has the focus and fringe attribute and 1315 randomly selected distant view images. 1000 images from each class of Set1 were set aside to extract features and train the SVM for the focus model while the leftover 522 close-up view and 315 distant view images were for testing. The remaining 1076 all focus close-up view and 1137 distant view images of the dataset make up Set2, where 800 images from each class were for training, and the others were the testing images. The scale model's CNN classifier for measuring object conceptual size used the training images of both Set1 and Set2. The experimentation were done in stages according to the focus and scale modeling.

4.1. Implementation Details

The key components of the focus model are the frequency transformation, feature extraction, codebook generation, and classification, while the scale model has two components, the spatial size evaluation based on object proposal methods, and the conceptual size evaluation using a fine-tuned CNN classifier.

4.1.1 Focus Model

With the exception of the classifier, where the SVM was used regardless of domain and features, the settings and details of the variants tested are as follows:

Domain	MWHFE	LBP	SURF(BoVW)	SURF(FV)
Original	51.52%	90.49%	87.65%	**94.38%**
DWT	75.99%	76.49%	87.83%	87.84%
NSCT-1	75.44%	74.24%	88.78%	89.84%
NSCT-2	74.74%	79.64%	86.19%	92.33%
NSCT-3	70.42%	85.73%	88.57%	93.76%
NSST-1	76.54%	70.08%	89.01%	89.70%
NSST-2	76.92%	73.67%	87.57%	88.53%
NSST-3	73.05%	80.24%	86.17%	91.04%

Table 1: Stage 1 classification results averaged over five cross-validation by random sub-sampling. (Gray cells show combinations used in Stage 2 of the experiment.)

(Domain) + (Features)	Accuracy (Stage1)	Accuracy (Stage2)	Difference
Original + LBP	90.92%	81.52%	-9.40%
NSCT-3 + LBP	83.87%	69.10%	-14.77%
NSCT-1 + SURF(BoVW)	90.92%	78.90%	-12.02%
Original + SURF(FV)	94.74%	82.00%	-12.74%
NSCT-2 + SURF(FV)	93.55%	79.31%	-14.24%
NSCT-3 + SURF(FV)	95.10%	80.14%	-14.96%
NSST-1 + SURF(FV)	91.16%	77.24%	-13.92%
NSST-3 + SURF(FV)	92.71%	78.14%	-14.57%

Table 2: Stage 2 classification results and accuracy difference between the performances in Stage 1 and Stage 2.

Frequency transformation. The DWT, NSCT, and NSST were used to obtain the high frequency details as explained in Section 3. Both NSCT and NSST are multi level decompositions, therefore, we implemented three levels in the experiments. For the NSCT, the decomposition scale directions used are $\{1, 2\}$ (NSCT-1), $\{1, 2, 8\}$ (NSCT-2), and $\{1, 2, 8, 16\}$ (NSCT-3), with the '9-7' pyramidal filter and 'pkva' ladder directional filter [42]. Whereas the NSST uses $\{1, 8\}$ (NSST-1), $\{1, 8, 16\}$ (NSST-2), and $\{1, 8, 16, 16\}$ (NSST-3) decomposition scale directions with the 'maxflat' pyramidal filter [25].

Feature Extraction. For the MWHFE, the histogram of energy is a column-wise summation of the high frequency map, which generates a 63-dimensions vector. Whereas, the LBP and SURF are natural image feature extractors that produce feature representation vectors of 10-dimensions and 64-dimensions respectively. Additionally, multiscale grid sampling is adopted in the SURF feature extraction. All three extractors were applied on the original RGB image, and the high frequency decompositions of DWT, NSCT, and NSST for a thorough comparison in the experiments.

Codebook Generation. The codebook generation step is only used in the classification pipeline using the SURF features. Both the vocabulary size for the BoVW and the number of clusters for the Gaussian Mixture Model (GMM) of the FV were set to 50.

Focus model \Proposal	Edge	Adobe	AdobeBING	RPN (ZF)	RPN (VGG16)
Original+LBP	86.41%	82.21%	89.38%	80.48%	87.17%
NSCT-3+LBP	80.97%	73.45%	88.38%	79.17%	84.97%
NSCT-1+SURF (BoVW)	87.59%	82.00%	92.48%	82.00%	88.83%
Original+SURF (FV)	88.07%	82.97%	92.21%	82.90%	88.69%
NSCT-2+SURF (FV)	87.03%	80.97%	91.79%	81.93%	88.14%
NSCT-3+SURF (FV)	88.07%	82.14%	**93.17%**	82.83%	89.45%
NSST-1+SURF (FV)	85.59%	79.93%	91.38%	81.17%	87.17%
NSST-3+SURF (FV)	86.90%	80.41%	92.28%	82.21%	88.48%

Table 3: Stage 3 classification results.

Focus model (+ AdobeBING)		Original + SURF (FV)		NSCT-3 + SURF (FV)	
Datasets	Actual \Predicted	Close-up view	Distant view	Close-up view	Distant view
Stage 1	Close-up view	498	24	492	30
	Distant view	20	295	11	304
Stage 2	Close-up view	592	206	550	248
	Distant view	55	597	40	612
Stage 3	Close-up view	735	63	782	16
	Distant view	110	542	83	569

Table 4: Confusion matrix of focus modeling methods in Stage 1, Stage 2, and combined with AdobeBING and fine-tuned CNN in Stage 3.

4.1.2 Scale Model

To evaluate the spatial size, the Edge Boxes (EB), Adobe Boxes (AB), BING refined by Adobe Boxes (AdobeBING), and the RPNs, ZF and VGG16 were implemented using the recommended parameters stated in their respective papers without any retraining of the used models. As detailed in Section 3.2.1, the spatial size is determined by selecting the top scoring box proposed by the methods, followed by the checking of the box area size. If the size is smaller than the threshold, the image is classified as a distant view, otherwise the area bounded will be used in the conceptual size classification. The threshold is fixed to be 20% of the image size. As for the conceptual size classification, the ImageNet pretrained AlexNet architecture [21] was chosen due to its simple and relatively small architecture, and its output layer is fine-tuned by re-mapping each of its 1000 object classes into either the big or small object class.

4.2. Stage 1: Focus Test

In the first stage of the experiments, the ability of the frequency domain and feature extraction techniques to distinguish the close-up image with focus and fringe attribute from the all focus distant view images were investigated. Therefore, only Set1 was used, where the 2000 training images were used to train the SVM classifier and the 522 close-up and 315 distant view testing images for models verification.

As detailed in Section 4.1, a total of four types of features (*i.e*. MWHFE, LBP, SURF (BoVW), and SURF (FV)) were tested on eight domains (*i.e*. the original image, DWT, NSCT-1, NSCT-2, NSCT-3, NSST-1, NSST-2, and NSST-3). The repeated random sub-sampling validation method was performed with five repetitions and the averaged performance are shown in Table 1.

We can see that the LBP, SURF(BoVW) and SURF(FV) perform well with accuracies above 80%, where the SURF(FV) applied on the original image is the best with the average accuracy of 94.38%. The MWHFE is the poorest performing very likely due to the insufficiency of representation by solely relying on the spatial summation of high frequency signals, as compared to the higher level representation provided by SURF.

4.3. Stage 2: Mixed Data Test

To verify the robustness of the focus modeling methods, the testing images in Set2, which contain close-up view images that have no focus and fringe attribute, are combined with the testing images of Set1. Therefore, the test data in this stage consist 798 close-up view and 652 distant view images in total.

According to the results of stage 1, eight combinations were selected for the experiment. The selection is made based on two criteria, (1) classification accuracy is above 90%, or (2) the number of distant views misclassified as close-up view is among the least. The second criterion is introduced to identify a model that has a better potential for improvement in overall performance when the scale model refines the results. (Further details in Section 4.4). The grayed cells of Table 1 shows the selected models. Table 2

Datasets	Approach / Actual \Predicted	[39] Close-up view	[39] Distant view	Proposed Close-up view	Proposed Distant view
Stage 1	Close-up view	354	168	492	30
	Distant view	30	285	11	304
Stage 2	Close-up view	395	403	550	248
	Distant view	63	589	40	612
Stage 3	Close-up view	683	115	782	16
	Distant view	117	535	83	569

Table 5: Confusion matrix of [39] and our proposed method in Stage 1, Stage 2, and Stage 3.

lists their results of testing on Set1[2] and Set1+Set2. A large drop, 13.33% on average, can be seen in the performance of all combinations which is contributed by the addition of the all focused close-up view images of Set2 into the test. Thus, it supports the necessity of a scale model for a better classification.

4.4. Stage 3: Scale Test

Stage 3 is dependent on the results produced by the combinations in Stage 2. This is because the scale model serves as the secondary classifier to address images that do not fit into the focus model. In other words, the scale model is used mainly to pick out the close-up view images that were misclassified by the focus model due to the lack of focus and fringe attributes.

Hence, from the results of Stage 2, all images that were classified as distant view by the focus models, regardless if they are correct or otherwise, are used for this Stage 3 experiment. As described in Section 3.2.1, five object bounding box proposal methods were tested for the spatial size evaluation with a fine-tuned CNN model for conceptual size classification. The fine-tuning of the CNN model uses the training set images of both Set1 and Set2, that are a total of 3600 images.

Table 3 shows the classification results where the best performing combination is the NSCT-3 domain using SURF features with FV codebook as the focus model, and the AdobeBING with CNN classifier as the scale model, at 93.17% accuracy. This is a different outcome comparing to the observation in Stage 1, where the best performing model in Stage 1 uses the original image for focus modeling instead of NSCT-3. Table 4 shows the confusion matrices of both models for each experiment stage for comparison.

Notably, using the original image in the focus model contributes to more misclassified distant view in Stage 1 and Stage 2 of the experiment as shown in the dark gray cells of Table 4. This is the decisive element in determining the best method because the scale model only handles misclassified close-up views, while the incorrectly classified distant views remains. For that reason, the NSCT-3 domain model is favored even though it performs worse in Stage 2, because most of them were rectified in Stage 3 as shown in the light gray cells of Table 4.

[2]All methods in Table 2 are trained and tested on the same image split for comparison, hence shows a different percentage than the average results in Table 1

4.5. Comparison with State-of-the-art:

Based on above experiments and analyses, NSCT-3 + SURF(FV) and AdobeBING + CNN are selected as the final framework and compared to the proposal by [39] which uses DWT + MWHFE and EdgeBoxes + CNN. By applying [39] onto this newly created dataset, its overall accuracy is 84%, while our framework achieves 93.17% with a statistically significant improvement of 9.17%. Table 5 shows detailed comparisons at each stage as highlighted by the light and dark gray cells where our new method always outperforms [39]. In the focus test, the NSCT-3 used in our proposal produces more sophisticated high frequency signals that preserves object appearance, whereas the DWT loses much of the details. In scale test, we found that the bounding box proposed by AdobeBING is more precise, while EdgeBoxes proposed small boxes that classifies these images as distant view, subsequently degrading the overall framework's performance. For more detailed comparisons, please refer the Section 1 in the supplementary materia.

5. Verifying the Correlation between Scenery Quality and View Ratio

Having establish a dependable framework, we investigate real travel photos and assess the potential of measuring the scenery quality of tourist spots based on view preference. The spots were selected based on three criteria: 1) Popularity: Recommended by top search engines - *National Geographic*, *Travel + Leisure* and *TripAdvisor*; 2) Objectivity: Having at least 4000 votes for each spot; 3) Generality: Evenly located in Asia, Europe, America and Oceania. After filtering, the distribution of hundreds of suitable candidates (the green curve) is shown in Fig. 5 that plots scenery quality rating (provided by TripAdvisor) against the number of spots having that rating. Based on those available locations, 12 spots, 10 high rating and 2 low rating, were

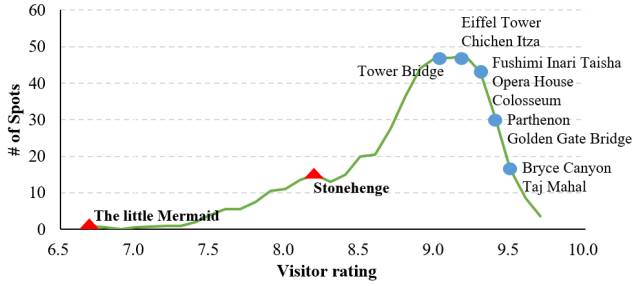


Figure 5: The positions of 12 selected spots on the number distribution of hundreds suitable spots at varied scenery quality ratings.

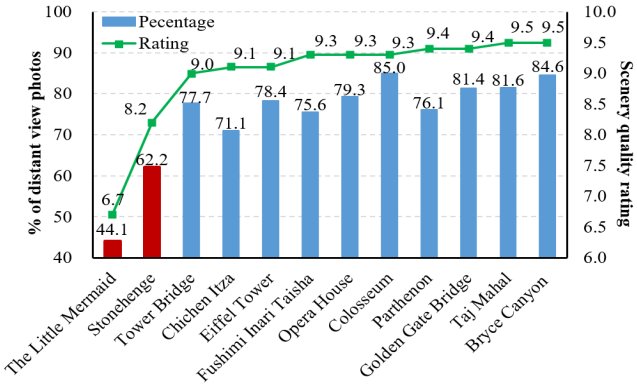


Figure 6: The scenery quality ratings from *TripAdvisor.com* and the proportion of distant view photos of 12 selected spots.

randomly selected as shown in the map of Fig. 1, for verification. More details can be found in the Section 2 of the supplementary material.

Then 2000 geo-tagged photos from each of these spots were downloaded from Flickr according to its geo-location and a specified radius (*e.g.* we used a circle zone, center at the *Eiffel Tower* with radius of 350 meters). Therefore, 24,000 photos in total were collected from 12 spots, which were categorized into the distant/panoramic view and close-up/local view images using our proposed framework. Figure 6 plots the proportion of distant/panoramic view (bar) and tourist rating (curve) of each spot. As per our hypothesis, there is a notable trend in the ratio of distant view to close-up view images. For the 10 high rating spots (above 9.0), there is an approximate ratio of $8:2$ or $7:3$, which is not apparent for low rating spots like The Little Mermaid and Stonehenge.

The proportion of distant view photos and the user ratings show a strong correlation with Person coefficient 0.956, which is also the first discovery of tourists' preference of a travel spot reflected in their photo taking habits. This statistic is a testament that the preference of the image viewpoints (visual attention) is able to bring forward the subconscious emotions induced by a tourist spot. Even so, this finding can in fact be explained by an assortment of human visual psychology literatures about the broaden-and-build theory [16, 33, 34].

For these reasons, we are convinced both psychologically and statistically that the emotional state brought by tourist spots with higher scenery quality is echoed in the photos taken by the tourists. Hence, we propose the ratio of distant to close-up view photos as a supplementary measure of travel locations with good outdoor scenery.

6. Conclusion

In this paper, we transformed the scenery quality evaluation problem into viewpoint classification task based on inspiration from the psychology theory of broaden-and-build. We investigated various computational algorithms to develop a framework that closely emulates the HVS and using this framework, we statistically analyzed travel photos from SNSs and found a strong correlation between the ratio of distant and close-up view photos with the rating of travel locations from TripAdvisor. Based on this statistical findings, and supported by psychological literatures, we proposed a distant to close-up view photo ratio as a supplementary scenery quality measure for outdoor travel locations that can potentially be developed into a full-fledged travel spot recommendation system.

Acknowledgments

This research is supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant FP070-2015A from the Ministry of Education Malaysia, Postgraduate Research Grant (PG002-2016A) from University of Malaya, and the Titan Z used was donated by NVIDIA Corporation. This work is also supported by the JSPS Grants-in-Aid for Scientific Research C (No. 15K00236).

References

[1] A. Barasch, K. Diehl, J. Silverman, and G. Zauberman. Photographic memory: The effects of volitional photo taking on memory for visual and auditory aspects of an experience. *Psychol. Sci.*, 2017. 1

[2] M. R. Basso, B. K. Schefft, M. D. Ris, and W. N. Dember. Mood and global-local visual processing. *J. Int. Neuropsych. Soc.*, 1996. 2

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 4

[4] I. R. Brilhante, J. A. Macedo, F. M. Nardini, R. Perego, and C. Renso. On planning sightseeing tours with tripbuilder. *Inf. Process. Manag.*, 2015. 1

[5] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014. 4

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop*, 2004. 4

[7] A. L. Da Cunha, J. Zhou, and M. N. Do. The nonsubsampled contourlet transform: theory, design, and applications. *IEEE TIP*, 2006. 3

[8] D. Derryberry and M. A. Reed. Anxiety and attentional focusing: Trait, state and hemispheric influences. *J. Individ. Differ.*, 1998. 2

[9] K. Diehl, G. Zauberman, and A. Barasch. How taking photos increases enjoyment of experiences. *Journal of personality and social psychology*, 2016. 1

[10] M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE TIP*, 2005. 3

[11] S. Duncan and L. F. Barrett. Affect is a form of cognition: A neurobiological analysis. *Cogn. Emot.*, 2007. 3

[12] G. Easley, D. Labate, and W.-Q. Lim. Sparse directional image representations using the discrete shearlet transform. *Appl. and Comput. Harmon. Anal.*, 2008. 3

[13] Z. Fang, Z. Cao, Y. Xiao, L. Zhu, and J. Yuan. Adobe boxes: Locating object proposals using object adobes. *IEEE TIP*, 2016. 4

[14] M. J. Fenske and J. E. Raymond. Affective influences of selective attention. *Curr. Dir. in Psychol. Sci.*, 2006. 3

[15] B. Fowler. 20 photos that will make you want to travel to budapest. www.businessinsider.my/20-photos-that-will-make-you-want-to-travel-to-budapest-2015-10/?op=0&r=US&IR=T. 2

[16] B. L. Fredrickson and C. Branigan. Positive emotions broaden the scope of attention and thought-action repertoires. *Cogn. Emot.*, 2005. 1, 2, 8

[17] K. Gasper and G. L. Clore. Attending to the big picture: Mood and global versus local processing of visual information. *Psychol. Sci.*, 2002. 2

[18] D. Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou. A survey on algorithmic approaches for solving tourist trip design problems. *J. Heuristics*, 2014. 1

[19] K. Guo and D. Labate. Optimally sparse multidimensional representation using shearlets. *SIAM J. Math. Anal.*, 2007. 3

[20] K. Hasegawa, Q. Ma, and M. Yoshikawa. Trip tweets search by considering spatio-temporal continuity of user behavior. In *DEXA*, 2012. 1

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 6

[22] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera. Personalized tour recommendation based on user interests and points of interest visit durations. In *IJCAI*, 2015. 1

[23] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan. Discovering areas of interest with geo-tagged images and check-ins. In *ACM-MM*. ACM, 2012. 1

[24] J. P. Lucas, N. Luz, M. N. Moreno, R. Anacleto, A. A. Figueiredo, and C. Martins. A hybrid recommendation approach for a tourism system. *Expert Syst. Appl.*, 2013. 1

[25] A.-U. Moonon and J. Hu. Multi-focus image fusion based on nsct and nsst. *Sens. and Imaging*, 2015. 5

[26] C. Morton. 20 photos that will make you want to visit antartica. www.cntraveler.com/gallery/20-photos-that-will-make-you-want-to-visit-antarctica/20. 2

[27] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cogn. Psychol.*, 1977. 2

[28] P. M. Niedenthal and S. Kitayama. *The heart's eye: Emotional influences in perception and attention*. Academic Press, 2013. 3

[29] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 2002. 4

[30] G. Pajares and J. M. De La Cruz. A wavelet-based image fusion tutorial. *Pattern Recognit.*, 2004. 3

[31] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 4

[32] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 4

[33] E. A. Phelps, S. Ling, and M. Carrasco. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychol. Sci.*, 2006. 1, 3, 8

[34] G. Pourtois, A. Schettino, and P. Vuilleumier. Brain mechanisms for emotional influences on perception and attention: what is magic and what is not. *Biol. Psychol.*, 2013. 1, 3, 8

[35] D. Quercia, R. Schifanella, and L. M. Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *ACM-HT*. ACM, 2014. 1

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4

[37] S. Schmalbruch. 33 photos that will make you want to travel to switzerland. https://www.msn.com/en-us/travel/tripideas/33-photos-that-will-make-you-want-to-travel-to-switzerland/ss-AAmIuVX. 2

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[39] S. Tong, Y. P. Loh, X. Liang, and T. Kumada. Visual attention inspired distant view and close-up view classification. In *ICIP*, 2016. 2, 3, 7, 8

[40] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE TPAMI*, 2002. 3

[41] J. K. Tsotsos. *A computational perspective on visual attention*. MIT Press, 2011. 2, 3

[42] Y. Yang, S. Tong, S. Huang, and P. Lin. Multifocus image fusion based on nsct and focused area detection. *IEEE Sens. J.*, 2015. 5

[43] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 4

[44] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, 2009. 1

[45] C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa. Anaba: An obscure sightseeing spots discovering system. In *ICME*, 2014. 3

[46] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 4