<table>
<thead>
<tr>
<th>Title</th>
<th>Sightseeing value estimation by analyzing geosocial images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Shen, Yizhu; Ge, Min; Zhuang, Chenyi; Ma, Qiang</td>
</tr>
<tr>
<td>Citation</td>
<td>2016 IEEE Second International Conference on Multimedia Big Data (BigMM) (2016): 117-124</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2016-08-16</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/2433/217605">http://hdl.handle.net/2433/217605</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.</td>
</tr>
<tr>
<td>Type</td>
<td>Conference Paper</td>
</tr>
<tr>
<td>Textversion</td>
<td>author</td>
</tr>
</tbody>
</table>

Kyoto University
Abstract—Recommending points of interests (POIs) is drawing more attention to meet the growing demand for tours. A POI’s quality (sightseeing value) estimation is one of the important challenges. In contrast to conventional studies of ranking POIs based on user behavior analysis, in this paper, we propose methods of quality estimation by analyzing geo-social images. Our approach estimates the sightseeing value from two aspects: (1) nature value; and (2) culture value. For the nature value, we extract image features that are related to favorable human perception to verify whether a POI would meet tourists’ psychological requirements. Three criteria, coherence, imageability and visual-scale, are defined accordingly. For the culture value, we recognize the cultural elements (i.e., architectures) included in a POI. In the experiments, by applying our methods on the real discovered POIs, we present the effect of our approach.

Index Terms—Points of Interests, Sightseeing value, Geosocial image, human perception, image processing

I. INTRODUCTION

Nowadays, travel is drawing more attention in people’s daily life. Benefiting from Social Networking Service (SNS) and advances in mobile devices, people can share their experiences on the Internet during travel. The vital information it contains provides researchers with excellent opportunities for discovering and ranking points of interests (POIs). For instance, GPS traces [1], images [2], check-ins [3], and tweets [4] are treated as different kinds of user votes to help gather tourism knowledge. Among them, an important challenge is how to evaluate the quality of a POI.

Although a lot of related work, such as [1] [3], have been done for POI recommendation, much is still unexplored. Based on a survey by Zheng et al. [5], the growing geo-referenced and community-contributed media resources have generated huge amounts of detailed location and event tags, covering not only popular landmarks but also obscure ones. As shown in Figure 1, we can divide POIs into four quadrants on the basis of two dimensions, i.e., “quality” and “popularity” [7].

Located in the quadrant with high sightseeing quality but low popularity, an obscure sightseeing location is a considerable choice for in-depth travel to not only enjoy the beautiful scenery but also experience local culture, especially for the repeat tourists who have already visited the most famous places there. In some senses, such kind of locations may be potential valuable sightseeing resources needed for developments and promotions. However, because obscure locations always have not enough visits or votes on the Internet, the conventional authority based analysis, which is used to recommend popular POIs, is not useful. In [6] [7], Zhuang et al. have proposed some methods to discover and rank the obscure locations that are not so well known. However, their methods still rely on the analysis of few users’ behaviors and the type of scenery objects (cherry blossom and maples are used as examples in their work), which make their solution not so flexible.

In this paper, by analyzing geo-social images, we propose a general approach to estimate the quality of both popular and obscure sightseeing spots. When people experience a landscape, there exists a process called human perception in which information is derived through senses, organized and interpreted [8]. In this way, a mental model [9] is devised that the human perception is influenced from the following three aspects:

1) the biological factors generated by evolutionary theory,
2) the cultural factors depended on cultural background, and
3) the individual factors resulted from individual differences in personality traits.

In accordance with this mental model, it is easy to know that cultural factors and individual factors vary from person to person, while the biological factors can be treated as cross-cultural commonalities for human perception of landscape. Therefore, we focus on the criteria served by the biological factors, which interpret the landscape from physical level to psychological level. By introducing the criteria, i.e., coherence, complexity,
disturbance, stewardship, imageability, visual-scale, naturals, historicity and ephemera, defined in environment psychology [10], we calculate image features as indicators to perform the quality estimation. Because these criteria are interrelated and interact on each other, our approach mainly focuses on the three key criteria: coherence, imageability, visual-scale and historicity. The first three criteria are related to CV (nature value i.e., sightseeing quality estimation from the nature perspective), while the fourth one views sightseeing spots from the angle of NV (culture value, i.e., the sightseeing value from the cultural perspective).

To the best of our knowledge, it is the first effort of investigating into estimating sightseeing value utilizing environmental psychology. To summarize, we make the following major contributions:

- **Content based methods for estimating sightseeing spots from the nature aspect:** By introducing the qualitative nature criteria defined in environmental psychology, we quantize three of them (i.e., coherence, imageability and visual-scale) for POI's NV estimation. To extract the indicators for the quantization, we devise several new algorithms to calculate the visual features from geo-social images taken in the target POI.

- **A time-based Analysis:** Because of the seasonal issues, a time series based analysis is further made to obtain the dynamic evaluation results for ranking POI candidates, based on which we can recommend different spots to users based on the season they are planning to visit.

- **A content based method for estimating sightseeing spots from the culture aspect:** Different from the human-based culture factors mentioned previously, there the culture refers to the inherent value held by the spot, which means we only estimate the culture in an objective way without considering the culture background for various tourists. Since some POIs contain several artificial elements (e.g., architectures), a heuristic method is proposed to measure the CV.

II. RELATED WORK

In this section, we first present the conventional related work on ranking POIs. Then, several studies on human perception for landscape environment will be introduced followed by several related work using image analysis. At last, some culture value related work will also be discussed.

**Ranking POIs** In the research into estimating sightseeing quality, a survey given by Luo et al. [11] shows that collections of geo-multimedia, which are a result of sightseeing experiences sharing by web communities, are widely used in trip recommendations. Ji et al. [12] modeled the relationships of scene/landmark and scene/authorship as a graph and adopted two popular link analysis methods, PageRank and HITS, to mine representative landmarks. Zheng et al. [13] aimed to mine interesting locations and classical travel sequences in a given geospatial region on the basis of multiple users' GPS trajectories. They first modeled multiple individuals' location histories with a tree-based hierarchical graph. Then, by using the graph, they proposed a HITS-based inference model that infers the interest of a location. In [1], the authors further developed a recommendation system. Instead of GPS traces, Liu et al. [3] proposed a joint authority analysis framework to discover areas of interest with geo-tagged images and check-ins. Hasegawa et al. [4] attempted to organize travel related tweets by considering the spatio-temporal continuity of user-behaviors during travel. By merging such fragmented tweets, users’ travel experiences can be detected.

In these work, GPS traces, images, check-ins, and tweets are treated as different kinds of user votes to help gather tourism knowledge. Authority based analysis, like “rank-by-count” and “rank-by-frequency” in a vote manner, is the basis for most of these trip recommendation research. However, for an obscure location, not enough visits or votes are generated on the Internet. The classical authority based analysis, which is always used to recommend popular sightseeing locations, is not suitable. Therefore, in our research, the human perception is introduced for solution.

**Human perception** There are many systematic analysis and studies on the human perception for landscape environment. Hartig [14] suggested that the movement into a landscape mainly resulted by evolutionary, sociocultural, and motivational forces. The existing of differences for natural landscape preference between user groups coming from different background is proved by the experiments done by Van den Berg et al. [15]. Ohta [16] proposed eleven cognitive criteria for evaluating natural landscapes and summarized a qualitative common structure for natural landscape cognition. M. Tveit et al. [10] gave an abstract framework for people’s interpretation for landscape from concept level to indicator level. In this framework, they proposed nine concepts for landscape.

These work propose the concepts and design disciplines for sightseeing value assessments and landscape restorations. In contrast, we propose novel quantitative analysis method for landscape assessment by exploiting geo-social images. To the best of our knowledge, our work is the first attempt at quantitative estimation of sightseeing values from the natural and cultural perspectives.

**Nature value** As in the image-processing field, several studies are trying to discover the relationships between images and human perception. Estimating aesthetic quality of a photo is highly related to our work. Tang et al. [17] extract both regional and global high-level features, and try to build connections between photo qualities and technical rules shared by photographers. Datta et al. [18] describe the aesthetic quality by selecting low-level features based on artistic intuitions. Furthermore, some more real scene depended features such as sky illumination [19] and landscape types [20] have also been considered for improving quality assessment.

In addition to these related work focusing on the evaluation for a single image rather than a real scene, the research that is quiet similar to ours is made by Berman et al. [21], who tried to uncover low-level image features that related to human perception of naturalness. Furthermore, Hunter et al. summarized the properties predicted to be important in usual environmental
theories and listed some measurable corresponded physical attributes for landscape preference in [22].

However, we argue that the low-level features implemented in these research work, such as color and spatial properties, are insufficient. To solve our problem, it is necessary to leverage such features into a higher level.

**Culture value** Culture is a very abstract concept including many sub-concepts such as art, design, history and so on and varies from different countries and regions. For example, Japanese culture is famous for special designed arts and crafts such as Niwa (Japanese traditional architecture), Ikebana (Growing flowers in a vase), Bonsai (Dwarf tree), Katana (Japanese sword), Kimono (Japanese traditional costume) and so on [23]. Our purpose is to evaluate the CV of a sightseeing spot using images. However, only a small part of these cultural elements frequently appear in the images taken by tourists. Traditional architecture is the most common cultural element that can be found in the images, which is a very import part of culture evaluation [24]. Traditional costume is another kind of cultural element that varies greatly from different culture [25] and contributes a lot to culture evaluation [26].

There are already some work on architecture parsing [27], detecting [28], style classification [29] and clothes style classification [30]. However, as far as we know, there is no work that combines the evaluation of sightseeing spot’s cultural value and the detection of these cultural elements. In this paper, we will build a bridge connecting these two parts.

### III. METHODOLOGY

In this section, we introduce our methods to estimate sightseeing values of a spot from natural and cultural perspectives. As shown in Figure 2, the input data of our methods are spots with geo-tagged images and the output are two sightseeing values from natural and cultural perspectives. Also, we can obtain sightseeing spots by applying clustering methods, such as DBSCAN to the geo-tagged images.

**A. NV Evaluation**

According to the landscape perception theory, the quality of landscape is affected by multiple factors. At first, we design corresponding image analysis method for each factor per each image. Then, we integrate these factors to obtain the NV of a spot with a set of images. In the final step, we arrange all the scores in a time series way, by which the seasonal issues are considered.

Based on Tveit’s study of environmental psychology [10], nine criteria should be considered for landscape assessments: coherence, complexity, disturbance, stewardship, image-ability, visual-scale, naturals, historicity and ephemera. To estimate the NV of a given spot with images, currently, we focus on the coherence, imageability and visual-scale, which are more realizable by utilizing image processing methods based on previous research.

1) **Coherence:** The coherence relates to the unity of a scene, enhanced by the degree of repeating patterns of color and texture [10]. Based on this definition, we consider color harmony, and repeated-pattern as detailed indicators to estimate the coherence of spots.

**a) Color Harmony:** Intuitively, colorful landscapes worth visiting. In this sense, we introduce color harmony as one indicator to estimate the NV based on coherence. The authors in [31] have proposed eight harmonious hue templates defined in a HSV space. As shown in Figure 3, each harmonious hue template contains a gray sector means a harmonic hue distributor for an image. All the areas and relative position relationships of sectors are fixed and only the rotation angle may change. An image whose hue distribution fits one of these templates can be regarded as having high color harmony.

Given an image, we use a harmony distance to calculate the difference between an image’s original hue distribution from the harmonious hue templates. The harmony distance with the most suitable template is defined as the color harmony score. We define each harmonious hue template $T_m$ as:

$$T_m = \{ (a_m, \omega_{m,k}); k = 1, \ldots, K_m \};$$

$$m \in \{ i, V, L, J, T, Y, X, I \}.$$
Fig. 4. The harmony distance calculated against each template for a given image. The top-left is the most matched template while the down-right is the worst one. 

$m$ means the 8 templates shown in Figure 3. The notation $K_m \in \{1, 2\}$ is the number of sectors in the $m$-th template and $\omega_m$ is the area of the $k$-th sector in the $m$-th template. $a_m$ is the rotation angle for the template $m$. The harmony distance from a given hue distribution to the $m$-th template is calculated by an appropriate $a_m$, which is introduced to minimize the distance as follows.

$$\arg \min_{a_m} \sum_{h} M(h) L_m(h, m)$$ \hspace{1cm} (2)

Where $h \in \{0, ..., 359\}$ is the index on each hue template. $M$ is a normalized hue distribution for an image and $L_m(h, m)$ is the loss function for $T_m$ in the hue position $h$. To define the loss function $L_m(h, m)$, we first introduce a Gaussian distribution $D(h, a_m, \omega_m, k)$, which is used to adjust the penalty of the loss function. The closer an index $h$ approaches to the boundaries of sector $k$ in template $m$, the larger the penalty will be.

$$D(h, a_m, \omega_m, k) = \frac{1}{\sqrt{\pi \omega_m}} \exp(-\frac{2(h-a_m)^2}{\omega_m})$$

Then, we use the loss function which used as color harmony $ch(i)$ for image $i$.

$$ch(i) = L_m(h, a_m) = \arg \max_{a_m} (D(h, a_m, \omega_m, k)) |k| + 1$$

when $\forall k \in \{1 \ldots K_m\}$, $|h - a_m| \geq \frac{\omega_m}{2}$ (i.e., $h$ is in the sector $k$); and

$$= \frac{\omega_m}{2\pi} \sum_{k \in \{1 \ldots K_m\}} (D(h^*, a_m, \omega_m, k) + 1)$$

$$+ \sum_{k \in \{1 \ldots K_m\}} (D(h, a_m, \omega_m, k_i) + 1)$$

when $\exists k \in \{1 \ldots K_m\}$, $|h - a_m| < \frac{\omega_m}{2}$ (i.e., $h$ is out of the sector $k$).

Figure 4 shows our calculation results using this algorithm. The template with the lowest harmony distance is considered as the most matched one for the given image and the distance value is used as the color harmony based score.

b) Repeated Pattern: If a landscape contains blocks with same or repeated patterns, the scenery is ordered and its coherence is considerably high. We consider the repeated pattern as the repeat or similar blocks showed in an image. In advance, we divide an image into blocks of $15 \times 15$ pixels and represent each block using a HSV space based histogram. We apply SOM (Self Organizing Map) [32] to cluster these blocks into $16$ ($4 \times 4$) groups.

To reveal the relative position of blocks, for any two groups $i$ and $j$, we use $n_{i,j}$ (see Figure 5(a)) to denote the occurrence times of the cases that group $j$ is in the adjacent positions of group $i$. The normalized $n_{i,j}$ can be seen as an occurrence probability of such a case.

Since all the blocks hold different saliency for perception, we calculate the average saliency score $a_{i,j}$ for groups $i$ and $j$ by using the saliency map method proposed by Harel et al. [33]. An example is shown in Figure 5(b). Based on the idea of weighted entropy [34], we get the repeated pattern score $rp(i)$ for image $i$ by following formula. Three examples are presented in Figure 5(c), by which we can observe that the more similar and ordered blocks an image has, the lower the repeated pattern score it holds. A low repeated pattern score means high coherence.

$$rp(i) = -\sum_{i=1}^{15} \sum_{j=1}^{15} a_{i,j} n_{i,j} \log(n_{i,j})$$ \hspace{1cm} (3)

We annotate $co(i)$ as the coherence for image $i$, and the $CO(s)$ as the coherence for spot $s$.

$$co(i) = \frac{ch(i) + rp(i)}{2}$$ \hspace{1cm} (4)

$$CO(s) = \frac{1}{n} \sum_{i=1}^{n} co(i) \ i \in \{i_{s1} \ldots i_{sn}\}$$ \hspace{1cm} (5)

2) Image-ability: The image-ability, which is defined as the strong visual image created by the landscape making people to have distinguishable and memorable experience, has a conceptional similarity with the photo quality assessment for image [10]. Therefore, we exploit photo quality assessment methods to estimate the imageability of a sightseeing spot. The idea is simple that if a spot has photos with high imageability, the sightseeing quality of that spot is reasonably high.

Currently, we use a machine learning method for this task. The database used for training are the images categorized as landscape in AVA dataset [35], which contains 250,000 images with aesthetic scores and semantic labels. We sort the images
with its average score and prorate the score with the value in the range from 1 to 5. Inspired by the work done in [17] [19] [20], we extract three low-level features to describe the whole image: the histogram of oriented gradients (HOG) [36], color moment (mean and standard deviation for RGB channel) [37] and local binary patterns (LBP) [38]. HOG is the feature widely used for object detection. LBP is the feature. LBP is found to be a powerful feature for texture classification and color moments, which colors distribution, is often used in image classification. About the training model, since a comparable output is expected, we use a cluster-weighted modeling (CWM) mentioned in [38] to do the regression and use the predicted value as the image-ability score, where the value range is from 1.0 to 6.0.

We annotate $im(i)$ as the Imagb-ability fot image $i$, and the $IM(s)$ as the Image-ability for spot $s$.

$$IM(s) = \frac{1}{n} \sum_{j=1}^{n} im(i) \quad i \in \{i_{s1}...i_{sn}\}$$  \hspace{1cm} (6)

3) Visual-scale: The visual-scale is defined as a perceptual unit that reflect the experience of landscape rooms, visibility and openness [10]. To calculate this criteria, we use the GIST based method introduced by Oliva et al. [39] to estimate the openness and depth using an image. The value range of both openness and depth is from 1 to 6. Here the openness refers to the view-shed size or the degree of occlusion of a landscape. The depth is more relevant to the max visual distance. Since both of openness and depth indicate the visual-scale of a landscape, we calculate these two value $op(i)$ and $dp(i)$ by using the model provided by Oliva [39] and take a average to calculate the visual-scale score for a spot.

We annotate $vi(i)$ as the Visual-scale fot image $i$, and the $VI(s)$ as the Visual-scale for spot $s$.

$$vi(i) = \frac{op(i) + dp(i)}{2}$$  \hspace{1cm} (7)

$$VI(s) = \frac{1}{n} \sum_{i=1}^{n} vi(i) \quad i \in \{i_{s1}...i_{sn}\}$$  \hspace{1cm} (8)

4) NV Calculation: We denote the input spot set as $S = \{s_{1}...s_{n}\}$. Each spot $s_{i}$ is represented by an image set. Because the NV of a spot is also affected by seasons, we divide the images into 12 months and try to implement these three evaluation method in a dynamic way. First, for each defined criteria (i.e., coherence, image-ability, visual scale), we construct three corresponding matrices: $M^{o}$, $M^{i}$ and $M^{v}$. Hereinafter, $M$ is denoted as one of the three matrices. $M_{i,j}$ is the average score of the target criteria for spot $s_{i}$ in month $j$ = {1...12}. Then, based on the $M$, three aspects are considered to evaluate $s_{i}$, overall level, durability and uniqueness. The overall level and durability is used to assign a high value for a spot with high and stable nature perception, which is perceived as a sightseeing spot suitable for large number of tourists. Besides, since people tend to pay more attention to find something special, the uniqueness, namely, assign a higher value while the other spots have relative low values for each month.

(1) The overall level.

$$Avg(s_{i}) = \frac{1}{12} \sum_{j=1}^{12} M_{i,j}; \quad i \in \{1..|S|\}$$  \hspace{1cm} (9)

(2) The durability.

$$Dub(s_{i}) = \sqrt{\frac{1}{12} \sum_{j=1}^{12} (M_{i,j} - \frac{1}{12} \sum_{j=1}^{12} M_{i,j})^{2}} \quad i \in \{1..|S|\}$$  \hspace{1cm} (10)

(3) The uniqueness.

$$Uni(s_{i}) = \frac{1}{12} \sum_{j=1}^{12} f(M_{i,j}) \quad i \in \{1..|S|\}$$  \hspace{1cm} (11)

Finally, the coherence, imageability and visual-scale based NV scores are calculated by mean, respectively.

$$NV(s_{i}) = \frac{1}{3} (Avg(s_{i}) + Dub(s_{i}) + Uni(s_{i}))$$  \hspace{1cm} (12)

B. CV Evaluation

The purpose of this part is to estimate the cultural value of sightseeing spots based on images taken by tourist. There are two challenges. The first one is that culture is a very abstract concept and hard to estimate. Our solution is to decompose the sightseeing spot into several objects and estimate the cultural value of each object in sightseeing spot. The second challenge is that all of our estimation is based on images, which means we have to choose the objects that appear commonly in images taken by tourists. We summarize five cultural elements which affect the cultural value obviously and appear commonly in photos. Table I shows the relationship among them. Architecture and its adornment and traditional costumes are very important cultural elements, which has been discussed in section II. In addition to these three elements, color preference is also a vital part of culture [40]. For example, the photos taken in New Yolk have a preference of blue gray color while people in Tokyo more like red and yellow colors. Besides, festivals and some cultural events reproduce the scene of traditional culture. If a cultural element never changed with time, we say it is static otherwise it is dynamic. If a cultural element could be defined based on only one object, we say it is object-dependent otherwise it is object-independent. For example, color preference never change with time but we can not define the color preference only by one object so it is static and object-independent. Conversely, traditional costume is worn by people and people can go to any sightseeing spots they like so it is object-dependent and dynamic. In this paper is our primary work and we only choose one of them, architectural style to estimate the cultural value of sightseeing spots, which at the same time includes several cultural elements like color.

Output of the last fully connected layer which contains 4096 also generates features of high quality. Here we use the In addition to good performance on classification, VGG net which achieves outstanding accuracy in images classification. VGG net [41] is a very well-known convolutional neural network the images belonging to the same category together. VGG natural scene and tourists. The purpose of this step is to gather spot contains different kinds of objects such as architecture, and bridges. One architecture category may contain many ob-

jects. We assume that the cultural value of sightseeing spots is increases the cultural value of sightseeing spots. Therefore, the

architecture of high culture quality and this object will be found in a cluster and U

classification confidence for each image. We assume that the regarded as of high culture quality. DPM gives a score of CV . The training set of DPM is consisted each other. For the form shown in image A and B respectively, M

AB is a set of matching points and scoreAB is a set of matching scores. Different size and resolution images will generate different number of SIFT points. Therefore we first calculate the average points that found in image A and B. If image A and B are depicting the same object, we can find a large number of matching points with low scores. In other words, the distance of image A and B is negatively related with the size of set M

AB and positively related with the score of each matching points. Therefore, the sum of scores of matching points is the numerator and the size of set M

AB is denominator. Finally, the distance formula is multiplied by the average SIFT points in image A and B to eliminate the effect of image size and resolution.

For images in the same cluster, if the distance of two images is smaller than a threshold, then we assume that there is an edge between them. Each image is assumed to be a vertex. Therefore we obtain an image graph for each cluster. An object is defined as a connected subgraph of each image graph. Images in a connected subgraph will be treated as they depicting the same object because they are similar enough to each other.

2) Architectural Object Separation: After clustering, each cluster will contain multiple objects. The images depicting the same object are similar to each other but obviously different from the images depicting other objects. SIFT [42] descriptor is highly suitable for this task. We define the distance between image A and B by the formula as follows

\[
D(A, B) = \frac{(n_A + n_B) \sum_{M_{AB}} \text{score}_{AB}}{2|M_{AB}|^2}
\]

where \(n_A\) and \(n_B\) denote the number of SIFT points found in image A and B respectively, \(M_{AB}\) is a set of matching points and \(\text{score}_{AB}\) is a set of matching scores. Different size and resolution images will generate different number of SIFT points. Therefore we first calculate the average points that found in image A and B. If image A and B are depicting the same object, we can find a large number of matching points with low scores. In other words, the distance of image A and B is negatively related with the size of set \(M_{AB}\) and positively related with the score of each matching points. Therefore, the sum of scores of matching points is the numerator and the size of set \(M_{AB}\) is denominator. Finally, the distance formula is multiplied by the average SIFT points in image A and B to eliminate the effect of image size and resolution.

For images in the same cluster, if the distance of two images is smaller than a threshold, then we assume that there is an edge between them. Each image is assumed to be a vertex. Therefore we obtain an image graph for each cluster. An object is defined as a connected subgraph of each image graph. Images in a connected subgraph will be treated as they depicting the same object because they are similar enough to each other.

3) CV Calculation: In this step, we train a Deformable Part Model (DPM) [43] to classify architectural objects and give a score of CV. The training set of DPM is consisted of several famous kinds of architectures which are widely regarded as of high culture quality. DPM gives a score of classification confidence for each image. We assume that the score of confidence is positively related with culture quality. The trained DPM model is conducted on each image of each object. If the max DPM score of an object is not larger than a threshold, we regard that this object is not related with architecture of high culture quality and this object will be omitted. The final CV score \(CV(s)\) of a sightseeing spot is given by the formula as follows.

\[
CV(s) = \sum_{C} \sum_{O} \frac{|U| \max(\sum_{O} D)}{|O|^2}
\]

where \(C\) denotes the set of clusters, \(O\) is the set of objects found in a cluster and \(U\) is a set of users who take the image of the object. \(D\) is a confidence vector given by DPM. Intuitively, for the \(D\), the more the objects with high DPM scores, the higher the \(CV(s)\) is. For the \(U\), if an object is of high CV, it should appear in many images taken by different users.

For implementation details, sightseeing spots will be divided into several clusters by VGG feature and k-means. Each cluster

---

### Table I

<table>
<thead>
<tr>
<th>Relationship among Cultural Elements</th>
<th>Object-dependent</th>
<th>Object-independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>Architecture, Adornment</td>
<td>Color Preference</td>
</tr>
<tr>
<td>Dynamic</td>
<td>Traditional Costume</td>
<td>Activity &amp; Event</td>
</tr>
</tbody>
</table>

Fig. 6. The harmony distance calculated against each template for a given image. The top-left is the most matched template while the down-right is the worst one.

The style of architectures varies greatly from different countries and regions. The existence of traditional architecture increases the cultural value of sightseeing spots. Therefore, the aim of this part is to detect architectures and classify them into different architectural style. There are many different kinds of architectures in one sightseeing spot such as towers, temples and bridges. One architecture category may contain many objects. We assume that the cultural value of sightseeing spots is positively related with the cultural value of each architectural object located in the landscape. Therefore generally, Figure 6 shows three steps for cultural value evaluation, category clustering, architectural object separation and cultural value evaluation for each architectural object.

1) Category Clustering: A set of images of a sightseeing spot contains different kinds of objects such as architecture, natural scene and tourists. The purpose of this step is to gather the images belonging to the same category together. VGG net [41] is a very well-known convolutional neural network which achieves outstanding accuracy in images classification. In addition to good performance on classification, VGG net also generates features of high quality. Here we use the output of the last fully connected layer which contains 4096 dimensions as feature descriptor and cluster them by k-means clustering method.

2) Architectural Object Separation: After clustering, each cluster will contain multiple objects. The images depicting the same object are similar to each other but obviously different from the images depicting other objects. SIFT [42] descriptor is highly suitable for this task. We define the distance between image A and B by the formula as follows

\[
D(A, B) = \frac{(n_A + n_B) \sum_{M_{AB}} \text{score}_{AB}}{2|M_{AB}|^2}
\]

where \(n_A\) and \(n_B\) denote the number of SIFT points found in image A and B respectively, \(M_{AB}\) is a set of matching points and \(\text{score}_{AB}\) is a set of matching scores. Different size and resolution images will generate different number of SIFT points. Therefore we first calculate the average points that found in image A and B. If image A and B are depicting the same object, we can find a large number of matching points with low scores. In other words, the distance of image A and B is negatively related with the size of set \(M_{AB}\) and positively related with the score of each matching points. Therefore, the sum of scores of matching points is the numerator and the size of set \(M_{AB}\) is denominator. Finally, the distance formula is multiplied by the average SIFT points in image A and B to eliminate the effect of image size and resolution.

For images in the same cluster, if the distance of two images is smaller than a threshold, then we assume that there is an edge between them. Each image is assumed to be a vertex. Therefore we obtain an image graph for each cluster. An object is defined as a connected subgraph of each image graph. Images in a connected subgraph will be treated as they depicting the same object because they are similar enough to each other.

3) CV Calculation: In this step, we train a Deformable Part Model (DPM) [43] to classify architectural objects and give a score of CV. The training set of DPM is consisted of several famous kinds of architectures which are widely regarded as of high culture quality. DPM gives a score of classification confidence for each image. We assume that the score of confidence is positively related with culture quality. The trained DPM model is conducted on each image of each object. If the max DPM score of an object is not larger than a threshold, we regard that this object is not related with architecture of high culture quality and this object will be omitted. The final CV score \(CV(s)\) of a sightseeing spot is given by the formula as follows.

\[
CV(s) = \sum_{C} \sum_{O} \frac{|U| \max(\sum_{O} D)}{|O|^2}
\]
IV. EXPERIMENTS

A. Overview

In this section, we investigate the effect of criteria calculation methods respectively and demonstrate the performance of our method for both NV and CV. Based on the algorithms introduced in Section III, for a certain spot, we make an estimation on all the images taken there and use the normalized Discounted Cumulative Gain (nDCG) [44] method to evaluate the results.

B. Dataset

We chose 14 sightseeing spots to be the experimental data: 7 spots in Kyoto, Japan and the other 7 spots in Suzhou, China. The spots in Kyoto are: Fushimi Inari Shrine, Kinkaku Temple, Ninna Temple, Tenryu Temple, Shisen-do, Hanami Street and Kyoto Station. The spots in Suzhou are: Zhuozhen Yuan, Tai Lake, Jinji Lake, Tiger Hill, Suzhou museum, Shantang Street and Guanqian Street. In this dataset, both high-quality spots which are abundant in natural elements and cultural elements (e.g. Kinkaku Temple, Shisen-do), and low-quality spots which mostly consist of modern architecture (e.g. Kyoto Station) have been considered to promise the unbiased of the experimental data.

Due to both of the confidence scores given by DPM and user preference are divided by the number of images, our method evaluates the cultural value of a sightseeing spot excluding the impact of the number of images. Generally, tourists prefer to going to famous sightseeing spot and taking more images compared with the sightseeing spots that are not so well-known. We can find more images related to popular spots than obscure spots. Therefore, in our dataset, popular sightseeing spots contain more testing images than others. However, we do not think more images means higher cultural value. Some obscure sightseeing spots of high culture quality are not crowded with tourists just for they have not been discovered. Thus we develop this method to evaluate cultural value of a sightseeing spot regardless of how popular it is.

C. Evaluation on NV

In accordance with our research, three criteria, i.e., coherence, image-ability and visual-scale, will be calculated and used for estimating NV. Hence, at first, we evaluate our methods to calculate these criteria.

For each criteria, we calculate the scores for all the images and take the average to describe the whole spot. The calculated criteria score for each spot are normalized and shown in Figure 7. Then we use the nDCG method to evaluate each criteria method with corresponding ground truth, which result is shown in Figure 8.
TABLE II
THE GROUND TRUTH: CULTURE AND NATURE SCORES OF EACH SPOT

<table>
<thead>
<tr>
<th>Avg. score</th>
<th>Tennryu Temple</th>
<th>Ninna Temple</th>
<th>Kinkaku Temple</th>
<th>Shisendo</th>
<th>Fushimi</th>
<th>Hanami Street</th>
<th>Kyoto Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>1k</td>
<td>1k</td>
<td>1k</td>
<td>1k</td>
<td>0.4k</td>
<td>1k</td>
<td>1k</td>
</tr>
<tr>
<td>Coherence</td>
<td>2.875</td>
<td>2.875</td>
<td>2.75</td>
<td>3.125</td>
<td>3.375</td>
<td>2.5</td>
<td>2.25</td>
</tr>
<tr>
<td>Image-ability</td>
<td>3.125</td>
<td>3.125</td>
<td>3.625</td>
<td>2.875</td>
<td>4.5</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>Visual-scale</td>
<td>3.375</td>
<td>3</td>
<td>3.25</td>
<td>2.625</td>
<td>2.375</td>
<td>2.75</td>
<td>2.375</td>
</tr>
<tr>
<td>Nature value</td>
<td>3.875</td>
<td>3.25</td>
<td>3.25</td>
<td>3.875</td>
<td>2.5</td>
<td>2.625</td>
<td>2.15</td>
</tr>
<tr>
<td>Culture value</td>
<td>2.875</td>
<td>3.75</td>
<td>4</td>
<td>3.375</td>
<td>4</td>
<td>3.25</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Avg. score</th>
<th>Tai Lake</th>
<th>Jinji Lake</th>
<th>Tiger Hill</th>
<th>Suzhou Museum</th>
<th>Humble Garden</th>
<th>Shantang Street</th>
<th>Guanqian Street</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers</td>
<td>1k</td>
<td>0.2k</td>
<td>1k</td>
<td>0.4k</td>
<td>1k</td>
<td>1k</td>
<td>0.3k</td>
</tr>
<tr>
<td>Coherence</td>
<td>2.625</td>
<td>4.25</td>
<td>3.75</td>
<td>2.25</td>
<td>3.125</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>Image-ability</td>
<td>2.875</td>
<td>3.75</td>
<td>3</td>
<td>3.125</td>
<td>3.5</td>
<td>3.75</td>
<td>2.25</td>
</tr>
<tr>
<td>Visual-scale</td>
<td>3.75</td>
<td>4.75</td>
<td>3.875</td>
<td>2.375</td>
<td>3.125</td>
<td>2.75</td>
<td>2</td>
</tr>
<tr>
<td>Nature value</td>
<td>3.875</td>
<td>3.625</td>
<td>4</td>
<td>2</td>
<td>3.75</td>
<td>2.625</td>
<td>1.25</td>
</tr>
<tr>
<td>Culture value</td>
<td>2.375</td>
<td>2.375</td>
<td>3.125</td>
<td>4</td>
<td>3.625</td>
<td>3.75</td>
<td>2.875</td>
</tr>
</tbody>
</table>

As introduced in Section III-A4, for NV estimation, we calculate an average value for all the images taken in each month by using the time-tag. Then we make a time-based nature evaluation by using these three criteria method in respective and combined way and then demonstrate the performance with nDCG method. Figure 9 shows the evaluation result.

1) Coherence: Based on our definition, the coherence mainly consists of two aspects: color harmony and repeated pattern.

According to the calculated result about coherence, it can be observed that the Fushimi Inari Shrine and Jinji Lake holds a relatively low coherence score, which refers to high coherence for visual perception. As introduced previously, the coherence is defined to be related to the unity of a scene, enhanced by the degree of repeating patterns of color and texture. It is easy to explain that since the images related to Fushimi Inari Shrine are mainly consist of toris that only have one color, red, a harmonic color tendency is generated for this spot. As a landside landscape where the scene is almost consists of the pure sky and pure lake, these simple repeated pattern in Jinji Lake brings people a high harmonic perception. The nDCG scores for coherence calculation is 0.9711.

The variation trend for color harmony and repeated pattern are shown in Figure 10 and Figure 11. The lower the color harmony scores and repeated pattern scores are, the higher the coherence held by the target spot.

According to the result for color harmony, it can be observed that the spots with many artificial architectures, i.e., Ninna Temple, Tenryuy Temple and humble administrator garden, are tend to hold a relatively smooth scores in the whole year. The reason is that since tourists pay more attention and take photos on the artificial architectures instead of nature elements, small impact is made by time. Based on the time analysis, the performance (nDCG) of nature evaluation implemented with only color harmony is shown in Figure 9, which is 0.974.

For the repeated pattern, the result shows that all the landscapes obtain smooth scores in a relatively fixed range except the Jinji Lake. As introduced previously, the pure sky and sure lake provides Jinji Lake a high coherence for human perception. Besides, the regular light events held in certain festivals also fluctuate on the repeated pattern score. By only
using repeated pattern, the performance (nDCG) of nature evaluation is 0.974.

2) Image-ability: By utilizing the method introduced in Section III-A2, the experimental result shows that our method gives high scores to Kinkaku Temple, Jinji Lake and Fushimi Inari Shrine, which matches the ground truth generated by subjects. However, low scores are calculated for Humble Administrator Garden while the ground truth for this spot is quite high. One considerable reason is that Humble Administrator Garden is a spot famous for its classical Chinese architecture, there are many landscape irrelevant images included in the experimental data, such as interior decoration and interior design. Recall that the definition of image-ability is strong visual image created by the landscape making people to have distinguishable and memorable experience. It pays more attention to outdoor aesthetic landscape more than indoor ones. The nDCG scores for image-ability calculation method is 0.963.

Based on the experimental results shown in Figure 12, it can be observed that the monthly distributions for image-ability of each spot seem to have no regular pattern. Since the photo quality is affected by many factors, such as composition, objects or even the focus of an image, it is difficult to distinguish whether a spot is beautiful or not just by considering photo quality. The nDCG result for image-ability is 0.967.

3) Visual-scale: With the methods proposed by Oliva [39], we extract GIST features from each images and use CWM to estimate the openness and depth. Then we calculate a harmonic mean to describe the overall visual-scale scores for spot.

According to the result shown in Figure 7, it can be seen that the calculated visual-scale scores for Jinji Lake and Tai Lake is explicitly different from other spots. The common feature for Jinji Lake and Tai Lake is that most of related images are about large lake scenery. Compared with the spots with a small space, this feature provides a stronger experience of wide-open appearance, which satisfies the definition of visual-scale in environment psychology. As shown in Figure 8, the nDCG scores of visual-scale calculation is 0.9803.

The experimental results for visual-scale shown in Figure 13 indicates that all the spots holds a smooth visual-scale score. The higher the visual-scale score is, the higher the visual-scale held by the target spot. It is easy to explain that the visual-scale is a fixed criteria for spot, which is invariable with time. The nDCG score for visual-scale based nature evaluation is 0.9756.

4) NV: For an overall NV estimation based on coherence, image-ability and visual-scale, we combine these three criteria by calculating the average of normalized score for each criteria. The performance of combined NV shown in Figure 9 is 0.9818, which is the highest performance we have got in the experiment.

5) Discussion: In our experiment, our method gives a lowest NV for Kyoto Station, which matches the round truth of nature perception. Intuitively, Kyoto Station is a spot with hardly any NV and CV. Figure 14 shows the representative photos for Kyoto Station. Most of photos are taken inside the station and filled with crowd. The result of the experiment result indicates that our method can with this classical case correctly and give low score, which is different from the other sightseeing spots.

However, we get the highest combined NV for Jinji Lake while the ground truth of Jinji Lake ranks behind Tiger Hill, Shisendo, Tenryu Temple and Tai Lake. According to interviews with subjects, one of the major reasons to assign a middle score to Jinji Lake is that although the major parts of the scene, i.e., sky and lake, belong to the nature element, the opposite buildings over lake still make a strong
artificial perception for the whole spot. Figure 15 shows the representative photos for Jinji Lake. However, our coherence based method gives a high evaluation to this spot because of its simple structure and clear blue color tendency, and the method for image-ability and visual-scale also gives a high score for its beautiful lake-view and broad field view, which leads a high combined NV for Jinji Lake.

We get a low NV on Humble Administrator Garden while its nature rank for ground truth is high. As explained previously in section IV-C2, since tourists tend to pay attention to interior decorations and interior designs rather than garden scenes, a large numbers of nature irrelevant photos are taken, which lead to a low score for both the visual-scale and image-ability, and a low coherence because of its complex indoor structure. The representative photos for Humble Administrator Garden are shown in Figure 16.

In short, our method tends to assign a high scores to spots with beautiful scene, wide fields of vision, obvious color tendencies or simple structures. Howeveras the Jinji Lake and Humble Administrator Garden case shows, it seems that this rule is not appropriate for all the high nature spots perceived by people. Besides that, a content bias for taken photos is another challenge for our method, which should be solved.

The nDCG scores show that most of the sightseeing spots are ranked correctly. Especially we have got the best performance when considering all of the three criteria. However, since each criteria has different degree correlation with NV, simply taking the average does not seem the best choice. As the future work, we will pay attention to the relationship between criteria and nature and give a appropriate coefficient for nature evaluation.

**D. Evaluation on CV**

To evaluate the performance of the cultural quality estimation, we first extract VGG features from images for each spot. Then, we use $k$-means to cluster images which is simple but well performed. Images of each sightseeing spot will be divided into 10 clusters. The threshold of step 2 in our method (see Section III-B) is set to be 100. We build an image graph for each cluster based on this threshold and objects are detected by looking for connected sub-graphs. In step 3, we download a training set from search engine, which contains 450 images and 15 classes. This image set is used for training DPM. The threshold of step 3 is set to be 0.

1) **Experimental Results:** As the result, for the sightseeing spots from Kyoto and Suzhou, the nDCG score is 0.9818.

2) **Discussion:** Our method gives Kinkaku Temple a rank of 4 while the true rank of Kinkaku Temple is 2. Kinkaku Temple is a very special sightseeing spot compared with other spots. It is famous for a golden temple and it appears in many images taken at this sightseeing spot. Recall that our method detects objects in sightseeing spots. In this case, our method can only discover one object, which is in a large number of images taken by different tourists. Although we take the number of photos taken by different users into consideration, the lack of objects still leads to a very bad rank of this spot. In our future work, we will make more efforts to find better methods to solve these exceptions.

Besides, there are still some problems need improve. For example, in addition to the scene images taken at a spot, there are also a lot of crowd images, food images, indoor images, and so on. Though some of them are filtered by Flickr’s key word based search, these noise images may affect our methods’ performance.

**V. CONCLUSIONS**

In this paper, we propose novel methods of sightseeing value assessment by analyzing geo-social images. We propose three criteria for nature value(NV) assessment: coherence, image-ability and visual-scale. We also propose the criteria for culture value(CV) assessment: architectural styles. Since the NV is affected by time, we also propose temporal analysis method of the NV. The experimental results demonstrate the performance of our methods.

As the future work, we will try to improve our criteria methods and find out the relationship between criteria and sightseeing value.

**ACKNOWLEDGMENT**

This work is partly supported by JSPS KAKENHI Grant Numbers 25700033, 15J01402 and 16K12532.