2013 IEEE 21st International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems

Camera Shooting Location Recommendations for Landmarks in Geo-Space

Ying Zhang
School of Computing
National University of Singapore
Email: yingz118@comp.nus.edu.sg

Roger Zimmermann
School of Computing and Interactive & Digital Media Institute
National University of Singapore
Email: rogerz@comp.nus.edu.sg

***Abstract*—Taking photos of landmarks is a favorite and popular way for travellers to keep memories of places they have visited. Community-contributed photo collections, such as on Flickr, provide us an opportunity to gain a more in-depth understanding of a landmark's visual appeal. While much current research is focusing on recommending *which* representative photos should be selected from such pervasive photo sources, our work aims to find out *where* a visitor can capture his or her own, beautiful and personal photo of a queried landmark. We believe that this aspect of helping users to take memorable photos has not been well studied. We propose a method to recommend a list of shooting locations that have the utmost potential to capture appealing photos for a landmark of interest. A Gaussian Mixture Model based clustering approach is applied to the camera locations from an existing photo repository, generating a set of regions each of which covers an area with sufficient semantics, *e.g.*, a route section. The scores and ranks among these camera locations are evaluated through multiple criteria, including their potential for better visual aesthetics, overall social attractiveness, popularity, etc. Additionally, we investigate the temporal characteristics of these locations by considering the spatio-temporal space. A number of different recommendations are generated from these results, such as the best camera positions at different times throughout a single day, or the best visiting time in the same spatial area. Subjective evaluation studies have been conducted, which indicate that our work can generate promising results.**

I. Introduction

Photo exploration has been one of the most direct ways for people to access visual information about landmark objects. The increasing popularity of media sharing services is creating an opportunity to better support users by sharing vast volumes of geo-referenced and community contributed resources, such as Flickr photos. Prospective tourists can benefit from such photo repositories by gaining a overall visual impression before visiting a landmark of interest.

There exist a number of studies that have attempted to understand tourist interests by leveraging such photo galleries. A popular approach is to return an image list as a result of a user query through either a text (*e.g.*, a landmark name) or a visual sample by similarity measurement. Some other methods extract representative photos of a queried object and present them with a user-friendly interface. All these techniques focus on *which* information can be retried from existing photo databases. However, they neglect an essential question: *how* does one obtain such information? For example, where should a nice snapshot be taken. In this work, our goal is to provide guidance to prospective visitors on where they may capture attractive landmark photos. Specifically, we recommend to visitors potentially good *camera locations* where they may be able to capture appealing landmark photos by themselves.

Several recent approaches have attempted to achieve similar goals by suggesting camera parameters, *e.g.*, rotation, vertical or horizontal shift, to users. However, this only works for slight and local adjustments within a continuous view from the current photographer's position, *i.e.*, an optimal subview from a view sequence at a fixed, given camera point. To the best of our knowledge, our method is the first to indicate potentially good camera locations for a landmark from a wide-area perspective. Additionally, our recommendations are performed not only in the spatial but also in the temporal domain, because for some landmarks certain time instances or durations are more likely to yield attractive landmark appearances, *e.g.*, at night-time for a building that has an elaborately lighted facade. For the remaining parts of this study we use the term *camera location* to indicate the general concept, but we may alternately use *camera spot* when emphasizing the spatio-temporal characteristics for both location and time. In our initial study the recommendations for landmark camera positions are based on the crowd-sourcing information of Flickr. However, other soucres could also be used.

The overall system architecture is presented in Fig. 1, with two main sub-components. Given a landmark name, our method initially maps a landmark name to its geo-coordinates using the GeoNames (www.geonames.org) gazetteer service. In the data preparation stage, photos are automatically crawled given the landmark positions and names using the Flickr API. For each photo, its images, attached meta-data (GPS location, time-stamp, user ID, etc.), user tags, and other information (*e.g.*, photo statistics) are retrieved. This crowd-sourced information provides the foundation for our method. Due to the semantic gap between the contextual information and the object itself, (*i.e.*, the landmark in our case), the retrieved photos represent various aspects and some of them may be irrelevant to the visual appearance of the landmark. Therefore, a tag based filter is applied to reduce noise and obtain a remaining set of photos that are highly landmark-relevant. We are especially interested in the ones that capture landmark external appearances, rather than the ones recording surrounding events or nearby people. To measure the quality of the location for its potential to serve as a base for an attractive landmark photo capture, a crucial indication is derived from existing photos' quality around this place. A recommended camera spot should be able to provide high aesthetics, *e.g.*,

1526-7539/13 $26.00 © 2013 IEEE
DOI 10.1109/MASCOTS.2013.25

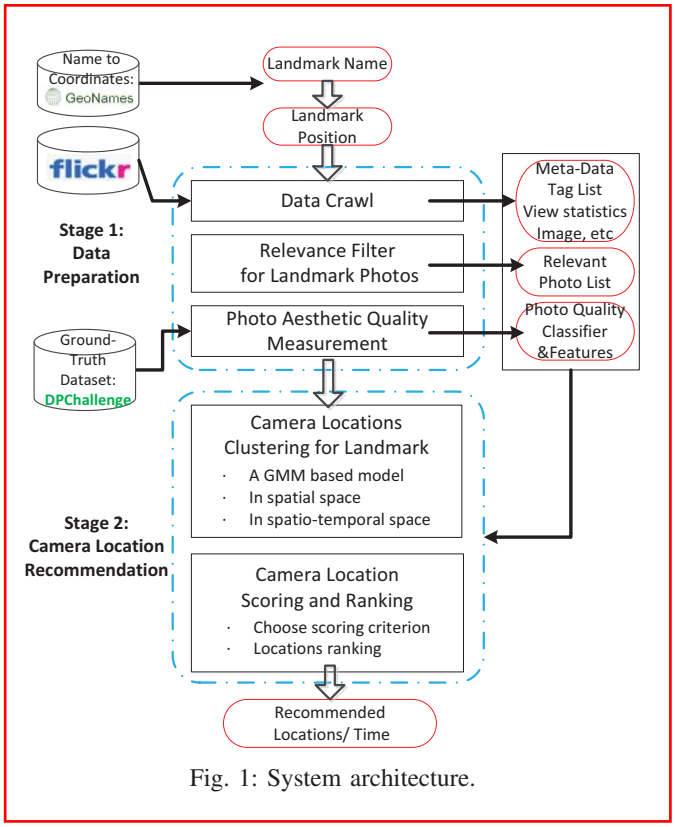


Fig. 1: System architecture.

picture composition, color distribution, etc. Hence we also incorporate a photo-quality measurement and obtain a classifier through Adaboost-based machine learning as well as several visual features to effectively distinguish aesthetics, supported by a ground-truth photoset from DPChallenge [1].

The second stage selects a list of locations or time-durations from where it is possible to capture attractive landmark photos. We observed that a good view location usually attracts crowds of visitors, however such popularity decreases gradually as the distance increases from the center. A Gaussian Mixture Model (GMM) is adopted to fit such a distribution pattern and divide photo locations into multiple partitions, either in the spatial (2D) or spatio-temporal (3D) domains. Each partition is a camera location candidate and we recommend the top k after measuring their potential to support a good landmark picture capture. The ranking mechanism measures the camera locations' overall quality by judging the quality of all the photos contained in the set. The measurement criteria include the photo aesthetics quality, the social attractiveness from photo viewing statistics, the overall popularity, the popularity consistency and the density distribution in both spatial and temporal dimensions. The final recommendation selects the top camera locations in the score lists, *e.g.*, the best camera positions in 2D, best visiting positions around a given time instance or the best visiting time at a give camera position (3D). The overall contributions of this paper are summarized as follows:

1) We propose a solution based on a Gaussian Mixture Model to partition camera locations into multiple regions. Experimental results indicate that most of these regions nicely outline the spatial characteristics of the surrounding areas of a landmark, *e.g.*, a route section or a small square. Such characteristics endows the clustering results with sufficient "semantics," which would be useful in many applications such as photo-indexing and exploration.
2) Camera location quality scoring is performed from the measurement of photos contained in a set, incorporating photo aesthetic factors, social attractiveness from viewing statistics, overall popularity, popularity consistency, spatial and temporal density. We successfully obtain an Adaboost based classifier to effectively and efficiently distinguish photos with appealing aesthetics from less-attractive ones, using a combination of three visual features and training with a real photo set. The overall error rate is approximately 22%, which is promising as we consider photos of any category and such features are providing a good foundation for general applications.
3) We further investigate the temporal dimensions of the photo set and conduct various recommendations, *e.g.*, best camera locations at a queried time period and optimal visiting time around a camera location.
4) A set of user studies are conducted and the experimental results imply that our framework can suggest good camera locations for a given landmark.

In the remaining parts of this paper, Sections II and III describe the related work and symbolic notations. Section IV elaborates on landmark-relevance filtering and Section V details the photo aesthetic quality measurements. Section VI introduces the camera location clustering and ranking. An experimental evaluation is presented in Section VII and conclusions are drawn in Section VIII.

II. RELATED WORK

A related research field is visual organization and exploration. For a user query, *i.e.*, a landmark, photos are traditionally searched by either their names or a sample photo by similarity matching from visual features, labeled tags or any other fusions. To diversify the searching results, some studies select representative photos for a queried scene [17], [11], [14]. For example, Kennedy *et al.* [11] cluster photos by investigating their visual features and select representative ones by several heuristics. As the visual diversity and representativeness matter their final goals the most, the connection between the locations and the visual features is weaken or even ignored. So some studies organize these landmark photos according to their location relations [8], [18], [4]. Epshtein *et al.* [8] adopt a hierarchical tree to organize the scene photos from an overview to details. At each level, representative images are selected for user to explore. Snavely *et al.* [18] organize visual repository in 3D space using full camera parameters, recovered by matching visual features between images. So tourists can be benefited by browsing these images according to the landmark 3D structure. Although the above studies integrate location factors into the problem, they still focus on the visual connectivity among photos that are geo-spatial adjacent but do not look into the location problem from a global perspective. It is observed that photos taken at locations that are apart from each other may also present similar views. Moreover, a detailed evaluation among different locations is

left blank. For real users, they still need to look through the whole architecture to find which views they like the most and need other supports to find out the camera location accordingly. In contrast, our work is a location driven study, *i.e.*, we investigate from the camera location perspective, looking for locations which have potentials to capture a nice landmark view. The final recommendations focus more on the quality and diversity of geo-locations, rather than the visual properties. We additionally examine the relations between location and time. There exist a few approaches incorporating time factor, however they usually focus on "event detection" [9], [6], [15] rather than bridging the gap between space and time. For example, Papadopoulos *et al.* [15] distinguish events-photos from others using visual features rather than photos' temporal information. In contrast, our work examines evaluations among multiple partitions in spatio-temporal space and suggests not only camera-positions but also time periods for an attractive capture.

To evaluate if a location has the potential for capturing appealing photos, one essential factor is the overall quality of the photos taken around that location. Quality is based mainly on such photos' visual appearance, *e.g.*, whether the distance to the landmark and the viewing-angles can generally create a well-balanced picture composition, or, if the surroundings provide a nice background and enhance the overall beauty. This inspired us to devise an overall photo quality metric. In this field, rule based feature extraction is a common approach. For example, Bhattacharya *et al.* [3] extract only the visual attention and check if it follows existing rules. Such features are more appropriate for an individual photo measurement but less so for a location measurement. A slight camera move can easily change the focus position and we require features that describe an overall visual impression [12], [5]. Cheng *et al.* [5] have proposed a method to automatically recommend user favorite photos from a wide-view or a continuous-view sequence. Photos are segmented into patches and modeled by the general distribution of each patch. Given a continuous view sequence a set of best camera parameters are suggested from this model. Similarly, Su *et al.* [20] designed a real-time view recommendation system by training an offline preference-aware aesthetic model. Yin *et al.* [22] extend the idea to mobile users. The ground truth input for these methods usually comes from photos with sufficiently high scores which however we found in our experiments to make up only a very small percentage of all photos. Another major limitation of these studies is that they target photos from a specific category, *e.g.*, landscape, which usually share similar visual patterns such as green grass at the bottom and blue sky at the top. However, for photos capturing the same landmark from different locations, the contained objects may vary a lot as they can be buildings, open space, garden, water, etc., which results in object shapes, textures, and colors hardly being the same. We desire general visual elements in highly-rated photos and in this way we can avoid modeling individual landmark locations.

III. Symbolic Notations

A landmark l is represented by a photo set P. For each photo $p \in P$, its camera location is determined by both longitude and latitude, $o = (lng, lat)$, and the capture time is indicated by the timestamp t. Multiple tags $x_1, \ldots, x_n$ are added by the photo's owner with user id u. Several weighting factors are used in this study and denoted with prefix w.

Fig. 2: Tags: (left) MBS, (middle) Singapore MBS, Marina, (right) Beer, Boatquay, Cars, F1, Formula 1, MBS, Singapore.

IV. Relevance Filter for Landmark Photos

Photo retrieval by landmark names can be noisy, owing to context ambiguities. A single tag could indicate multiple intentions and a tag list may present multiple facets of an object (this is sometimes also refered to as the *intention gap*). For instance, the landmark name may indicate that the photo captures the landmark itself, or events that happened nearby, or people walking around, or a number of other intentions. As the final goal of our work is to recommend potentially good camera locations to capture the landmark, it is essential and necessary to perform image filtering. Only photos that are highly related to the landmark itself, especially its exterior appearance will be retained for further processing. In our method, such landmark-relevance of each photo is determined by all attached user-labeled tags. One advantage of a context-based approach is its light-weight computational complexity and hence user-generated tags may partially bridge the gap between the content and its semantics.

The key idea of tag filtering relies on the observation that photos captured close to each other tend to have similar visual contents together with similar tags. Hence, tags occurring frequently among a set of spatially adjacent photos are more representative. Consider the three examples, all of the Marina Bay Sands (MBS) hotel complex, in Fig. 2. Tags of the right-most photo describe more a short-term event (F1, Formula 1) and personal activities (Beer, Boatquay, Cars) than the building itself (MBS, Singapore). So the tags with frequent occurrence indicate a stronger relationship to the landmark.

K-Means clustering is first applied to divide the locations of landmark photos into k groups $P_1, \ldots, P_k$ with tag sets $X_1, \ldots, X_k$. Tags in each group are analyzed and ranked by their occurrence frequency. For each tag x, three attributes are calculated: the inner frequency counting the total occurrences among the cluster: $f_{in}(x) = count(x, X_i)/|X_i|$ (where $count(a, A)$ counts the total occurrences of element a in set A), the external frequency of occurrences among other clusters $f_{ex}(x) = count(x, \bigcup_{j \neq i} X_j)/|\bigcup_{j \neq i} X_j|$ and n_x, the total number of distinct users using the tag x. The tag landmark-relevance is determined by these three factors under a TF-IDF combination, similar to the work by Ahern [2]. However our filter narrows the scope to within a single landmark: $r(x) = f_{in}(x) \cdot \frac{1}{f_{ex}(x)} \cdot n_x$. For a photo with multiple tags, its landmark-relevance depends on the summation of the individual tag relevances: $r(p) = \sum_{i=1}^{m} \{r_{x_i}\}$. The relevance threshold is selected to balance the precision and recall as will be illustrated by the experiments in Section VII-B.

V. Photo Aesthetic Quality Measurement

A camera location is closely related to the quality of the photos that have been taken nearby. With a higher percentage of high quality photos captured in the area, the location should have a greater opportunity to be recommended. We first investigate which visual aspects may determine a photo's quality by checking if some common features exist in highly rated images. Unlike existing studies that target images within a single category, *e.g.*, landscapes or beaches with sunrises – where visual patterns are straightforward such as the sun and the sea are usually separately by sky, or the blue sky appears above the green grass – we would like to allow images to be of any category. Hence we exploit a machine learning approach to explore the features that could be commonly utilized. A photo set with sufficient real user ratings is leveraged as the training data. Multiple intrinsic features such as color and texture are investigated and we describe our approach in the next paragraphs.

A. Color Features

We extract a color histogram to investigate the color distribution of photos with high or low ratings in either RGB or HSV color space. The histogram of each channel under each color space is calculated by partitioning each channel into 16 bins. For each image, this feature is finally characterized by a 48-dimensional vector.

To describe other color features we also extract the Color Moment [19], which is widely used in image classification and content based image retrievals. We choose three central moments of an image's color distribution: the mean, the standard deviation and the skewness. Hence, each image is presented by nine moments total with three moments for each color channel.

B. Texture Features

To model the texture characteristics, we adopt a well-known feature – histogram of oriented gradients (HOG) [7] – by counting the occurrences of gradient orientations in localized portions of an image. HOG is widely used in computer vision and image processing for the main purpose of object detection. A grayscale image is divided into four by four subregions, from each of which a local histogram of eight gradient directions over all pixels is counted. Finally we arrive at a 128-dimensional HOG feature vector.

We also adopt the Local Binary Patterns (LBP) [13] for texture modelling. LBP has been found to be a powerful feature for texture classification by encoding the neighboring eight pixels with respect to the pixel value at each location. Additionally, LBP improves detection if combined with HOG in some cases [21]. For each image, a 256-dimensional vector is extracted to represent LBP.

C. Spatial Distribution Features

Spatial organization indicates the relationship among all objects in an image and a balanced spatial organization over the whole picture plays an essential role in highly rated photos. We propose two features to represent such spatial distributions.

First we analyze edge distributions using Difference of Gaussian (DOG). Two Gaussian filters with $\sigma_1 = 5$ and $\sigma_2 = 2$ are applied to the monochrome image and pixels with positive values from the difference between these two Gaussian filters are detected as edges. The spatial distribution is characterized by partitioning the image into 4×4 subregions and counting the total number of pixels related to the edges.

Second we extract a salience map [10], a topographically arranged map representing the visual saliency of a corresponding visual scene. The most salient location in a visual scene would be a good candidate for attention selection. In our studies, we prefer that photos taken at a recommended camera location include the landmark in its salient regions. A salience map is extracted from each monochrome photo. Dividing the photo into 4×4 subregions, we count the total number of pixels whose salience values consist of the top 25% and reshape them into a 16-dimensional vector.

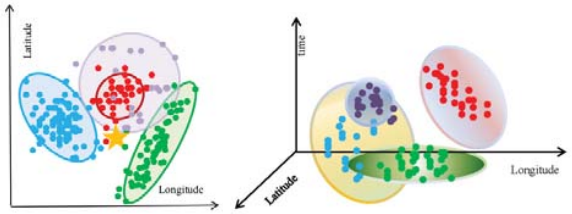
D. Feature Classification

We finally imported the above features into both SVM and Adaboost based classifiers. The features and classifier parameters were selected to achieve good classification results, with detailed procedures and outputs presented in the experiments Section VII-C. In the next step the photo aesthetics results are used to obtain camera location recommendations.

VI. Camera Location Recommendation

A. GMM-based Camera Location Clustering

A good camera location attracts additional photos to be captured in the area, however such popularity decreases gradually as the distance increases from the cluster center. This property suggests to describe the landmark camera location distribution with a Gaussian Mixture Model (GMM). Each Gaussian component represents the distribution of one camera location candidate.



(a) 2D (b) 3D

Fig. 3: Illustration of Gaussian Mixture Model (GMM) based camera spot clustering in 2D and 3D.

We denote a location with $o = (lng, lat)$. For a landmark we assume that there exist a total of K camera locations within its visibility area $\mathbb{G}$, which can be derived from the size of each landmark [23]. Then the spatial distribution of variable o is given by the GMM model:

$$P(o) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(o|k),$$

where $\mathcal{N}(x|k)$ is the density under the k^{th} Gaussian component and w_i is the prior probability of the k^{th} component with

$\sum_{i=1}^{K} w_i = 1$, $0 \leq w_i \leq 1$. $\mathcal{N}$ follows the distribution of a single Gaussian model:

$$\mathcal{N}(o|k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp -\frac{1}{2}(o-\mu_k)^T \Sigma_k^{-1}(o-\mu_k),$$

where μ_k and Σ_k denote the mean and covariance matrix of k^{th} component and d represents the data dimension. To fit a dataset containing N photos into a GMM distribution, we use the EM algorithm to determine all the parameters μ_k, Σ_k and w_k, $k = 1, \ldots, K$. In the E-step, the probability whether a sample belongs to the k^{th} Gaussian component is decided. The class assignment probability is calculated as below:

$$\begin{aligned} P(k|o) &= \frac{w_k \cdot \mathcal{N}(o|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} w_{k'} \mathcal{N}(o|\mu_{k'}, \Sigma_{k'})} \\ n_k &= \sum_{i=1}^{N} P(k|o_i) \end{aligned}$$

The M-step determines the model parameters:

$$\begin{aligned} \mu_k &= \frac{1}{n_k} \sum_{i=1}^{N} P(k|o_i) o_i \\ \Sigma_k &= \frac{1}{n_k} \sum_{i=1}^{N} P(k|o_i)(o_i - \mu_k)(o_i - \mu_k)^T \\ w_k &= \frac{\sum_{i=1}^{N} P(k|o_i)}{\sum_{k=1}^{K} \sum_{i=1}^{N} P(k|o_i)} \end{aligned}$$

These steps are repeated until they converge and at most K camera location candidates are obtained. The clustered components, if sufficient samples exist, have the potential to define an outline of the spatial characteristics within an area, such as a route segment nearby a landmark. This bestows the clustering with meaningful semantics.

B. Camera Location Recommendation

We propose to assess the quality of these camera areas according to the six criteria $s_1, \ldots, s_6$, listed below:

- **Photo quality** (s_1): The higher the aesthetics of a set of spatially nearby photos, the higher the probability that the area where these photos were taken is a promising place, *i.e.*, $s_1 = \frac{1}{N} \sum_{i=1}^{N} q(p_i)$, where $q(p)$ is the aesthetic quality according to Section V.

- **Total area popularity** (s_2): In general, a better camera location attracts more photo snapping visitors and such popularity gradually accumulates over time. The combined individual appearances of all photos form a global impression of the landmark, especially for some classic locations. As a single person may take multiple photos, we use an entropy based summation to reduce such impact. For the unique user set U assume that each user u_j has taken m_j photos. The photo popularity is calculated as a summation of individual contributions:

$$\begin{aligned} s_2 &= \sum_{j=1}^{|U|} (1 + \sum_{i=2}^{m_j} log(\frac{i}{i-1})) \\ &= \sum_{j=1}^{|U|} 1 + \sum_{j=1}^{|U|} (log \prod_{i=2}^{m_j} \frac{i}{i-1}) \\ &= |U| + \sum_{j=1}^{|U|} log(m_j) \end{aligned}$$

- **Area popularity consistency** (s_3): A good view location should attain a steady popularity over time. We use the standard derivation of the times when the photos were taken to measure such consistency. The larger the value, the higher the variability, hence $s_3 = \sqrt{\sum_{i=1}^{N}(t_i - t_0)^2}$, where $t_0 = \frac{1}{N} \sum_{i=1}^{N}(t_i)$.

- **Photo spatial density** (s_4): The spatial coverage of different view-locations varies. Camera locations that are far from a landmark usually occupy a larger spatial coverage area than the ones closely surrounding the landmark. A plausible reason is that the landmark usually is proportionally smaller in a scene if captured far away, so two photos could still be similarly-looking even though they are geo-spatially apart from each other. However the picture focus may be completely different if they are taken at the same distance but from different directions. So we measure the average number of photos over their spatial distribution. A higher photo density indicates a more promising place: $s_4 = N/\mathbb{G}(S)$.

- **Photo temporal density** (s_5): In any recommendation system, it is essential to consider "cold starts." In other words, the location quality should not entirely depend on the total popularity, otherwise it may miss newly-emerging areas. If a location has the potential for an appealing landmark view, it would gain popularity much quicker in the temporal dimension. So we investigate the photo density over time, $s_5 = N/(max(t) - min(t))$, where t is the photo capture time. Higher values indicate better locations. This way we capture newly emerging, but promising regions.

- **Social attractiveness** (s_6): Community-contributed resources are often collected on an open and free platform which connects media objects and people. Major photo-sharing websites record audience activities with regard to a media object, including a total count of views, a total count of favors, comments, ratings, etc. All these statistics are valuable sources to indicate the public affection of an object. Social attractiveness is a subjective assessment that indicates how a photo is liked by others [11], [22]. For a single photo, statistics are calculated as $a = (w_v \cdot v + w_f \cdot f) \cdot \mathfrak{t}$, where v and f denote the total count of views and favors, $\mathfrak{t}$ is an indicator for statistical confidence from post duration with normalization. Both v and f are cumulative factors and the longer the posted duration $\mathfrak{t}$, the higher confidence these statistics contribute to the photo's subjective quality. Not all the landmark photos have sufficient statistics to indicate their quality due to various reasons such as photo access restrictions. From

our observation, only 4% have more than 5 favors and 17.4% have more than 100 views. To remove low-level noise, we only incorporate the photos that have more than 50 views, $s_6 = \sum_{p_i \in P} a_i$.

After obtaining scores for all critera, we linearly combine them to compute the overall score of a single camera location, *i.e.*,

$$s = \sum_{i=1}^{6} w_i \cdot s_i, \quad (1)$$

where w_i represents the weight to adjust the contribution of each criterion to the overall score. The photo quality is additionally related to visibility, *i.e.*, whether the landmark can be captured without visual blocking. Some existing work [16] has indicated the possibility to derive such information using complicated geometrical computations, which need support from 3D maps but are beyond the scope of this study. In the future, we will explore more of such factors.

C. Spatio-temporal Camera Spot Recommendations

The discussion up to now has concentrated on the spatial relationship between the camera and landmark locations. However, in many cases, landmarks might be more appealing during certain time periods. For example, a building having an elaborately lighted facade, while interesting during day-time, may be most appealing during the night. Another landmark with beautiful landscaping would be better captured under good lighting conditions so that the picture can include surroundings for a better visual appearance. These scenarios have inspired us to incorporate an additional factor, time, into the solution. In this section, we investigate camera spots with different spatio-temporal features and present their semantics. The time scope is initially set within a single day, but this temporal scale is adjustable (*e.g.*, a month, a year) to generate various semantics.

Going forward, each photo has three attributes: latitude, longitude and timestamp. The timestamp should be converted to the time zone in which the landmark is located. GMM is applied to these 3D points, dividing the spatio-temporal space into multiple partitions. We employed a direct 3D clustering instead of further dividing the previous 2D clustering results due to two reasons. First, 3D clustering avoids having to set the same number of sub-clusters of previously generated clusters with different time spreads so it can better model the time continuity. Second, it avoids partitioning the data into groups with too few samples. Hence, photos that are both spatially and temporally close will be assigned together. For each generated camera spot (a GMM component), its recommendation score is measured by averaging qualities over all containing photos as in Eqn. 1. Next, we introduce several terms that will be used in this section:

SPATIAL COVERAGE: For a camera spot S, its spatial coverage $\mathbb{G}(S)$ is the main projection area of the Gaussian component to the spatial surface (*i.e.*, lng-lat dimensions). A regular Gaussian distribution covers the whole domain of definition with the highest density at the center and a gradual decrease as deviating from the core. To simplify the problem, we reduce the target spatial projection to be the area containing over 85% of its elements, which can be regarded as a confidence interval.

SPATIAL NEIGHBORS are defined as the camera spots whose spatial coverage have sufficient overlap, $\mathbb{G}(S_i) \cap \mathbb{G}(S_j) > Th_s$.

TEMPORAL BURST: A component S is a temporal burst if it has spatial neighbors and its quality outperforms all others, *i.e.*, $\forall j \neq i \in \mathbb{G}^\tau(i), s(S_i) > s(S_j)$, where $\mathbb{G}^\tau(i)$ contains all spatial neighborhoods of component S_i.

TEMPORAL COVERAGE is denoted as $\mathbb{T}(S_i)$, which is the Gaussian component's projection over time dimension.

TEMPORAL NEIGHBORS share sufficient overlap of their temporal coverages as $\mathbb{T}(S_i) \cap \mathbb{T}(S_j) > Th_t$.

SPATIAL BURST: A component S is a spatial burst if it has temporal neighbors and its quality outperforms all others, *i.e.*, $\forall j \neq i \in \mathbb{T}^\tau(i), s(S_i) > s(S_j)$, where $\mathbb{T}^\tau(i)$ contains all temporal neighborhoods of component S_i.

Our goal is to find all spatial and temporal bursts for a queried landmark as outlined in Algorithm 1. Initially a flag is set to true for each camera spot candidate (Line 2). For each candidate S with a true flag, all its neighbors (either spatial or temporal ones, Line 7) are collected. Next, they are ranked according to their quality scores and only the top one is labeled as a burst (Line 8). All others are flagged false to avoid checking in the future (Lines 9-11). Various types of applications can benefit from this, *e.g.*, to suggest the best photography time or location for a single landmark according to personalized requirements. It can also be helpful for trip planning among multiple landmarks, *e.g.*, when deciding a visiting order among these destinations in the temporal dimension as well as an optimal viewing location for each of them during the assigned time period.

Algorithm 1 Find temporal (spatial) bursts for a landmark.

Require: $\{S\}$: camera location candidates.
1: $B = \emptyset$: burst.
2: $\{flag_1, flag_2, ..., flag_n\}$: burst flag for each S, initally true
3: **for** $S_i \in \{S\}$ **do**
4: **if** $flag_i = false$ **then**
5: continue;
6: **end if**
7: $\mathbb{G}^\tau(i)$=findNeighbors($\{S\}, S_i$)
8: S_k= getTop (getSCORE($\mathbb{G}^\tau(i)$))
9: **for** $S_i \in \mathbb{G}^\tau(i), S_i \neq S_k$ **do**
10: $flag_i = false$
11: **end for**
12: $S_i \rightarrow B$
13: **end for**
14: Return B

VII. EXPERIMENTS

A. Data Setup

We crawled 52,264 photos from Flickr of 15 landmarks from 6 countries including the United States (Statue of Liberty, Washington Monument, Lincoln Memorial, Japantown, Coittower), France (Notre Dame Cathedral, Arc de Triomphe, Eiffel Tower), Singapore (Marina Bay Sands, Singapore Flyer, Merlion, Esplanade), Australia (Sydney Opera House), Japan (Tokyo Tower) and Italy (Colosseum). The photos are retrieved by their names and location constraints (within a circle centered at the landmark position and 3 kilometer radius). The

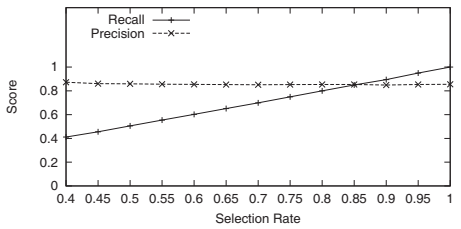


Fig. 4: Evaluation of the image filter results, the x-axis is the selection rate and the y-axis is the score.

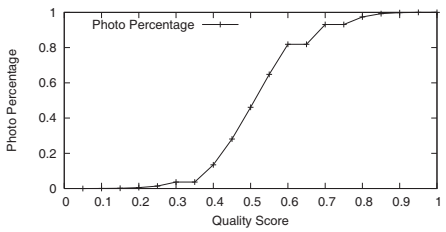


Fig. 5: Cumulative distribution function (CDF) of photo quality percentage.

queried landmark names are enriched from the knowledge on Wikipedia which provides popular alternate names if possible. *E.g.*, the Notre Dame Cathedral is enriched with Notre Dame de Paris and Notre Dame. The landmark geo-coordinates are obtained from GeoNames. A few gray images are removed as we need to extract color features later. The metadata attached to each photo is also retained, including timestamp, user id, latitude and longitude, etc.

B. Image Filtering

For each landmark, tag based filtering is applied to retain landmark-relevant photos. K-means clustering is used initially over all photo geo-coordinates. In our experiments we used $k = 10$, however this is adjustable to achieve best results. Tag ranking is conducted for each cluster, where the tags attached to all photos within a cluster compose one "internal" set and tags among all other $k - 1$ clusters constitute another "external" set. Both internal frequency, external frequency and the total count of unique users are computed. So for each photo, its relevance score is a summation over all its tags. Photos within the top m percentage of the score lists are selected as landmark-relevant, which was the input for the next step. We manually labeled photos over several landmarks and the experimental results are illustrated in Fig. 4. We choose the intersection point between recall and precision (85%) to filter irrelevant images. We also investigated if the total number of tags would affect the filter results and we found that, in our case, this factor did not help much to improve the performance.

C. Image Quality Measurement

1) Ground Truth Dataset: We accessed an open photo set [1] including 16,509 images and their meta-data (image ID, total number of ratings, rating values). Each of the images had at least 100 ratings from a public audience ranging from 1 to 10. To distinguish between images with good or plain aesthetic quality, we sorted the images by their mean rating and selected the top 10% and bottom 10% images as aesthetics learning input.

2) Feature Selection and Classifier Training: To find proper features to distinguish images with good versus plain aesthetic quality, we tested eight individual features: HSV histogram, RGB histogram, HSV color moments, RGB color moments, Histogram of Oriented Gradients (HOG), Difference of Gaussian (DOG), Local Binary Pattern (LBP) and Salience Map. Each image was first resized to 500 by 333 pixels. Both Adaboost and SVM based classifiers were applied to the image set and the overall classification error rates are illustrated in Table I. The smaller the value the better the feature performs. From the statistics we found that salience map and DOG perform the worst with approximately a half correct classification rate. Color features rank second with histograms performing slightly better than the color moments. Feature LPB performs the best.

For the classifiers, Adaboost performs better than the SVMs so we further used three Adaboost classifiers to train the data of combination features from the top features. The classification error rates in percentage with three combinations are presented in Table II. The best results achieved an error rate of 22.54% with the combination of LBP, HOG and HSV color moments, using the Real Adaboost classifier. Though slightly lower than a few other state-of-art outcomes, our results are reasonable as other work considered a special set of images, *e.g.*, a natural landscape image set, while our dataset contains various contents categories.

TABLE II: Classification error rate (%) of feature combos.

Classifier	LBP, HSV--histogram	LBP,HSV--moments	LBP, HoG, HSV--histogram
Modest Adaboost	23.557	23.516	23.091
Real Adaboost	23.735	23.003	22.543
Gentle Adaboost	23.676	24.162	22.891

3) Quality Measurement on Flickr Dataset: We selected both the features and the classifier from the experiments above and applied them on our dataset. The image quality is generated as a probability between 0 (low) and 1 (high). The cumulative distribution of the total percentage of photos as the maximal score increased is illustrated in Fig. 5. The distribution implies that the image quality is spread over the whole range which is reasonable as Flickr images are usually from the public with diverse photography skills.

D. Viewing Statistics

The viewing statistics for each public Flickr photo are available and we collected both the number of total viewings and the number of favors for each photo in our dataset. Table III presents the Flickr photos' distribution under different viewing statistic ranges from which we can see that the total number of favors is quite small with a majority having no favors ($83.47\% = 33,971/40,699$). These statistics are taken as one of the metrics for location quality.

E. Camera Location Recommendation

1) Recommendation in Spatial Space: K-Means clustering was applied to initialize the center position and covariance matrix of Gaussian components, reducing the iterations. The number of components was set to ten, which is adjustable to

TABLE I: Classification error rate (%) of eight individual image features.

Classifier	HSV histogram	RGB histogram	HSV color moments	RGB color moments	HOG	LBP	DOG	Salience
Gentle Adaboost	37.806	38.515	38.309	41.159	38.165	23.667	43.986	46.495
Real Adaboost	37.398	38.545	38.731	40.770	38.553	24.046	44.076	46.588
Modest Adaboost	36.935	38.458	38.836	42.511	37.698	24.369	43.108	45.696
SVM(linear)	49.169	47.173	50.351	49.873	49.458	49.936	49.904	49.618
SVM(RBF)	50	49.984	49.904	49.904	44.264	49.904	49.904	49.904
SVM(tuning)	38.123	38.442	35.441	38.634	33.780	25.891	40.294	46.360

(a) Marina Bay Sands (MBS) (b) Sample photos (c) Singapore Flyer

Fig. 6: An overall distribution of Gaussian components and the positions of their containing photos for landmarks(a) MBS (c) Singapore Flyer and (b) their sample photos.

TABLE III: Viewing statistics (number of total views and favors) for Flickr photos.

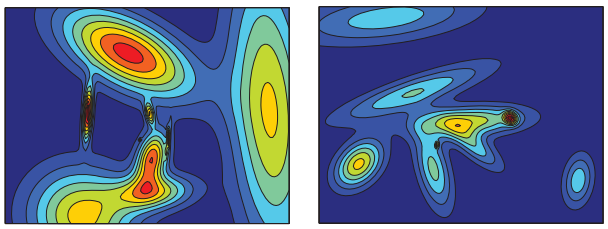
# of Views	# of Photos	# of Favors	# of Photos
=1	6,777	=0	33,971
1-10	11,531	1	5,200
10-50	15,310	5	659
50-100	5,808	10	440
100-500	715	20	283
500-1k	465	20-100	141
1k-5k	52	100-1,000	5
>=5,000	41	>=1,000	0

achieve different partition granularities. We may obtain at most 10 Gaussian clusters, depending on the data distribution. Each photo is assigned to the cluster with the maximum posterior probability for observation, weighted by the component probabilities. Figs. 6(a) and (c) are two examples, presenting the Gaussian component distribution and their containing photos for Marina Bay Sands (MBS) and Singapore Flyer (Flyer). The plot is on top of a map with landmarks indicated by a green star. Gaussian components are outlined by ellipses, whose positions and sizes are determined by Gaussian center and covariance matrixes, and different components are distinguished by color. Note that the real distribution would cover the whole space and each ellipse is only the area covering the sample majority. From the results we found that some of these components well outline the shapes of their nearby routes. For example, components 6 and 9 in Fig. 6(a) shape the Esplanade Bridge and Art Science Museum, separately. Components 10 and 1 in Fig. 6(c) occupy Raffles Avenue and the Flyer Pte building. It is observed that some components contain a few noisy samples which failed to be filtered and we will investigate such scenarios in the future. As these two landmarks are located in the same area, we found that they partially share some components, however with different components sizes. *E.g.*, component 4 of the Flyer is dividing components 6 and 10 of MBS. A plausible explanation is that the landmark sizes are different. The Flyer, as a relatively smaller object compared to MBS, if captured from a long distance, differs in appearance in a photo not much from a big object like MBS which leads the two components of MBS to merge. We measured the location quality using Eqn. 1. The camera location score is the average score over all its containing photos. The higher the component quality, the higher a rank that photo is assigned. The ranking results for MBS and Flyer are:

```
Marina Bay Sands:(top)6,8,5,7,10,1,9,4,2,3
Singapore Flyer :(top)1,7,10,9,4,6,5,3,2,8
```

To visually illustrate the locations, we selected sample photos from each camera location (GMM component) in Fig. 6(b), labeled from 1 to 10. Due to space limits, rankings with different weights are not presented but it is interesting to see that some components are top with all weights on one factor but at the bottom when using the other. *E.g.*, component 9 of MBS scores high on aesthetics however at the bottom on social attractiveness. This may be due to the aesthetic critera usually incorporating various factors however a viewer may be prone to favor a photo when the picture-focus clearly matches the target (the landmark in our case). This is inspiring us to investigate features emphasizing the landmark in the future. To be better aware of the overall quality distribution, contour maps are sketched for the above two examples in Fig. 7, which are geographically matched to the maps in Fig. 6.

2) Recommendation in Spatio-temporal Space: To further investigate the effects from temporal characteristics on the camera location quality, we expanded the Gaussian Mixture Model from 2D space to 3D, looking for components that are



(a) MBS (b) $Flyer$

Fig. 7: Contour maps for (a) MBS and (b) the Flyer.

dense in both spatial and temporal dimensions. The time scope was a single day.

To understand the relations among these spatio-temporal components, we recommended both (1) spatial and (2) temporal bursts. Fig. 8 shows samples for MBS. Around 7 PM, component 6 receives a higher score as a spatial burst than 4. A sample photo of component 6 shows a front and paranomic view of MBS, however component 4 captures only an upwards and partial view. Similarly, at 4 PM, the Esplanade Bridge (component 9) captured a general impression and is a better view platform than the water side (component 8), which focuses the left part of the landmark rather providing a balanced view. For temporal bursts, component 9 is more popular in the afternoon rather than the late evening and component 4 earns a higher score in the morning than in the evening.

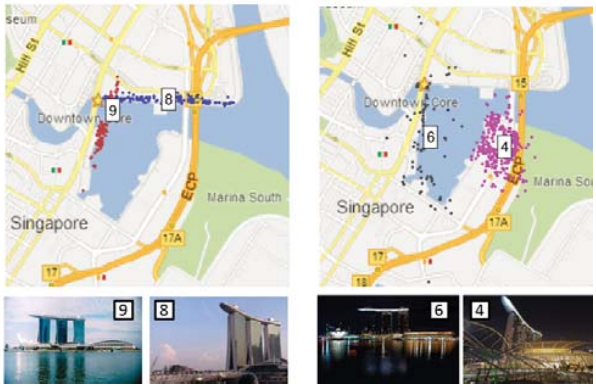


Fig. 8: Spatial/Temporal bursts for MBS, and sample photos. Camera spots 8 and 9 are at 4 PM. Camera spots 6 and 4 are at 7 PM. Camera spot 6 shares partial spatial coverage with 9.

Another example for the Flyer is shown in Fig. 9. Around evening time (7 PM), camera location 7 is preferred over 4, probably owing to its easiness to capture a complete landmark shape, and the surrounding water is a beauty bonus. For temporal bursts, 1 PM of camera location 3 outperforms its spatial neighbor 7 at 7 PM. A possible explanation could be the bright light conditions outlining the Flyer more clearly together with the blue sky and green trees, all of which help to enrich the photo colors.

3) User Study Evaluation and Discussion: As there exist neither other related work with the exact same research goals nor a set of standard critera in this field, we conducted a user study to evaluate our framework. A set of results for two landmarks in Singapore are selected for evaluation. 15 people participated in the survey, all of whom are living in Singapore and are familiar with the selected two landmarks.

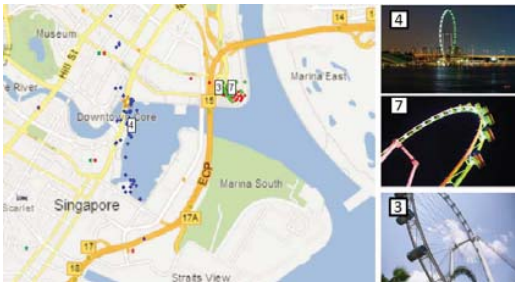


Fig. 9: Spatial/Temporal bursts for Flyer and sample photos. Camera spots 4 and 7 occur around 7 PM. Camera spot 3 share similar spatial coverage with spot 7.

The survey had five parts: 1) Some general questions about the work motivation; 2) 10 pairs of image-sets were given, each set corresponding to a recommended location in space and each pair between the top and the bottom selections. Participants were asked to compare the quality between each pair; 3) 6 pairs of images from spatio-temporal recommendations, each a spatial/temporal burst, were compared; 4) We partitioned the ranking components into 5 parts (P1 to P5) from worst to best, selected images for each part and asked the participants to rank them from least-favorite to most favorite; 5) The last part compared the top 5 locations from our algorithm with two other methods: one a random location selection (baseline) and the other from a related study [23]. That work aimed to generate video summarizations of landmarks in geo-space. The proposal is to use both the distance and the viewing direction from the camera to the landmark to evaluate video salience. The viewing direction was excluded in the experiments as we could always obtain an image with the landmark in the center. We applied spatial clustering to photo locations and selected photos from the top 5 clusters according to their average distances. Participants were asked to score each method's results from 1 (poor) to 5 (best). All participants were trained to judge the landmark capture performance rather than other objects, and to disregard the diversity of photo appearances.

Except for one user, all others have searched online photo repositories for landmark images before traveling. 13 out of 15 agreed that the location recommendation is useful during travel planning. Fig. 10(a) illustrates all participants' expectations for each pair of photo sets. The shorter the distance, the better the results our method generated. From the figure we can observe that the overall results are satisfactory and all of them cross the mid-point of 0.5. We looked into the samples with a lower score, *e.g.*, our algorithm scores 7 (sample in Fig. 6) higher than 5 for the Flyer, however there still exist users liking 5 more. A plausible reason is that the capture angle around location 5 could include the surrounding buildings, trees and water for a nice photo composition. However the views from 7 have many buildings surrounding, which may weaken the clarity of the Flyer itself. Fig. 10(b) suggests an overall agreement for our generated spatial or temporal bursts. But some expectated distances were a little bit larger than the distance in Fig. 10(a). We checked into the results and found that people show less preference for temporal bursts among spatial neighbors which implies that people are less sensitive to temporal factors. Fig. 10(c) shows the participants' evaluations from the best locations to the worst which generally follow our algorithmic results, except for the

set P4 of MBS (Samples 4 and 9 in Fig. 6). This set is expected to be better than both sets 2 and 3 (Sample photos 1,5,7,10), which indicates a general preference for a complete over a partial view. Fig. 10(d) compares the top selection using our method with both MVS and random-baseline. The statistics illustrate that our results outperform the others. However, random-selections are sometimes even better than MVS, since the closeby distance cannot capture a complete landmark.

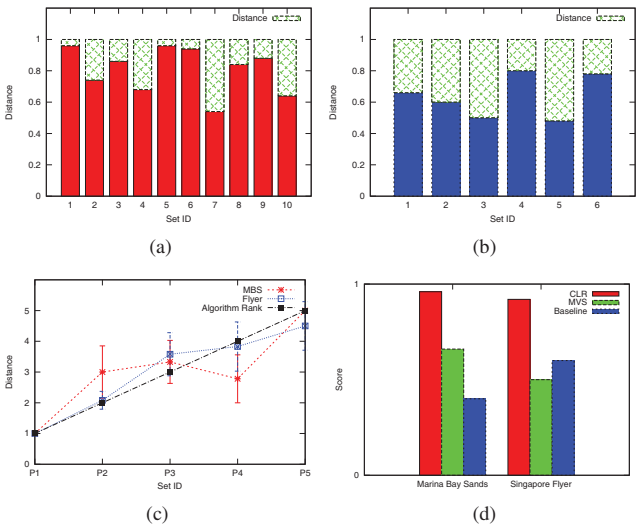


Fig. 10: (a) All participants' expectations for each pair of a photo set under spatial (2D) clustering. (b) Similarly, all participants' expectations under spatio-temporal (3D) clustering. (c) Ranking evaluation. (d) Comparison among our algorithm, MVS and a random baseline.

VIII. CONCLUSIONS

We presented a system for recommending camera locations with a high potential from which to shoot good photos of a landmark. A Gaussian Mixture Model based clustering is proposed to partition photo locations into multiple components. Experiments show that many components outline the shapes of surrounded routes well. Camera location quality judgement was converted to the measurement of its containing photos, incorporating aesthetic factors, social attractiveness from viewing statistics, overall popularity and consistency, spatial and temporal densities. We successfully obtained an efficient Adaboost classifier using three visual features and trained with a real photo set. The overall error rate of around 22% is promising to distinguish photo quality, given general a photo domain. We further investigated the temporal characteristics of locations by conduction a clustering in the spatial-temporal domain. Different recommendations can be generated with these results, such as the best camera positions throughout a day, or the best visiting time around a spatial area. Subjective studies were conducted to evaluate the framework and the experimental results support that our framework suggests good camera shooting locations for a given landmark.

ACKNOWLEDGEMENT

This research has been supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

REFERENCES

[1] DPChallenge Dataset. ritendra.weebly.com/aesthetics-datasets.html.

[2] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. In *ACM/IEEE-CS Joint Conference on Digital Libraries*, 2007.

[3] S. Bhattacharya, R. Sukthankar, and M. Shah. A Framework for Photo-Quality Assessment and Enhancement Based on Visual Aesthetics. In *ACM International Conference on Multimedia*, 2010.

[4] A. Brahmachari and S. Sarkar. View Clustering of Wide-Baseline N-views for Photo Tourism. In *Conference on Graphics, Patterns and Images*, 2011.

[5] B. Cheng, B. Ni, S. Yan, and Q. Tian. Learning to Photograph. In *ACM International Conference on Multimedia*, 2010.

[6] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal Event Clustering for Digital Photo Collections. In *ACM International Conference on Multimedia*, 2003.

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[8] B. Epshtein, E. Ofek, Y. Wexler, and P. Zhang. Hierarchical Photo Organization using Geo-Relevance. In *ACM International Symposium on Advances in Geographic Information Systems*, 2007.

[9] K. Gavric, D. Culibrk, P. Lugonja, M. Mirkovic, and V. Crnojevic. Detecting Attractive Locations and Tourists' Dynamics using Geo-Referenced Images. In *International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services*, 2011.

[10] J. Harel, C. Koch, and P. Perona. Graph-based Visual Saliency. In *Advances in Neural Information Processing Systems*, 2007.

[11] L. S. Kennedy and M. Naaman. Generating Diverse and Representative Image Search Results for Landmarks. In *International Conference on World Wide Web*, 2008.

[12] C. Li and T. Chen. Aesthetic Visual Quality Assessment of Paintings. *IEEE Journal of Selected Topics in Signal Processing*, 2009.

[13] T. Ojala, M. Pietikäinen, and D. Harwood. A Comparative Study of Texture Measures with Classification Based on Featured Distributions. *Pattern Recognition*, 1996.

[14] S. Palmer, E. Rosch, and P. Chase. Canonical Perspective and the Perception of Objects. *Attention and Performance IX*, 1981.

[15] S. Papadopoulos, C. Zigkolis, S. Kapiris, Y. Kompatsiaris, and A. Vakali. ClustTour: City Exploration by Use of Hybrid Photo Clustering. In *ACM International Conference on Multimedia*, 2010.

[16] Z. Shen, S. Arslan Ay, S. H. Kim, and R. Zimmermann. Automatic Tag Generation and Ranking for Sensor-Rich Outdoor Videos. In *ACM International Conference on Multimedia*, 2011.

[17] I. Simon, N. Snavely, and S. Seitz. Scene Summarization for Online Image Collections. In *IEEE International Conference on Computer Vision*, 2007.

[18] N. Snavely, S. M. Seitz, and R. Szeliski. Photo Tourism: Exploring Photo Collections in 3D. In *ACM International Conference and Exhibition on Computer Graphics and Interactive Techniques*, 2006.

[19] M. Stricker and M. Orengo. Similarity of Color Images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 1995.

[20] H. H. Su, T. W. Chen, C. C. Kao, W. Hsu, and S. Y. Chien. Preference-Aware View Recommendation System for Scenic Photos Based on Bag-of-Aesthetics-Preserving Features. In *IEEE Transactions on Multimedia*, 2012.

[21] X. Wang, T. Han, and S. Yan. An HOG-LBP Human Detector with Partial Occlusion Handling. In *IEEE International Conference on Computer Vision*, 2009.

[22] W. Yin, T. Mei, and C. W. Chen. Crowdsourced Learning to Photograph via Mobile Devices. In *IEEE International Conference on Multimedia and Expo*, 2012.

[23] Y. Zhang, H. Ma, and R. Zimmermann. Dynamic Multi-Video Summarization of Sensor-Rich Videos in Geo-Space. In *International Conference on Multimedia Modeling*, 2013.