Rare is Interesting: Connecting Spatio-Temporal Behavior Patterns with Subjective Image Appeal

Gökhan Yildirim
School of Computer and Communication Sciences
EPFL, Switzerland
gokhan.yildirim@epfl.ch

Sabine Süsstrunk
School of Computer and Communication Sciences
EPFL, Switzerland
sabine.susstrunk@epfl.ch

ABSTRACT

We analyze behavior patterns and photographic habits of the Nokia Mobile Data Challenge (NMDC) participants using GPS and time-stamp data. We show that these patterns and habits can be used to estimate image appeal ratings of geotagged Flickr images.

In order to do this, we summarize the behavior patterns of the individual NMDC participants into rare and repeating events using GPS coordinates and time stamps. We then retrieve, based on both the time and location information from these events, geotagged images and their "view" and "favorite" counts from Flickr. The appeal of an image is calculated as the ratio of favorite count to view count. We analyze how rare and repeating events are related to the appeal of the downloaded Flickr images and find that image appeal ratings are higher for events when the NMDC participants also took pictures and also higher for rare events. We thus design new event-based features to rate and rank the geotagged Flickr images. We measure the ranking performance of our algorithm by using the Flickr appeal ratings as ground truth. We show that our event-based features outperform visual-only features, which were previously used in image appeal ratings, and obtain a Spearman's correlation coefficient of 0.47.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Experimentation, Human Factors, Measurement, Verification

Keywords

Geo-tagging; image appeal rating; spatio-temporal behavior patterns; time- and location-based events

1. INTRODUCTION

Whether we like a photograph or not is highly subjective and depends on personal preferences. We can consider an image as appealing for various reasons: we find it interesting, aesthetically beautiful, or emotionally touching. An image that is appealing to a person can be mundane to another. Here, we aim to show that we can infer the appeal of a photograph by analyzing the daily life patterns of people.

The recent developments in smart phones allow us to take high-resolution photographs with accurate time and Global Positioning System (GPS) information. Moreover, social networks enable people to upload millions of photographs and share them with the community on crowd-sourcing web sites such as Flickr[1], where they will be rated by many people. We can thus investigate the relation between people's behavior patterns and the photographic appeal using smart phone data and crowd-sourcing websites. However, there is a practical disconnect between these data sources, which is a challenge to overcome.

On one hand, geographical data, namely geotags, can be successfully mined through mobile devices. These geotags later can be used to estimate people's daily life patterns. However, only a portion of the photographs collected with these mobile devices are shared on web sites, either because of privacy concerns or because many photos just never "leave" the phone. On the other hand, we can find very large numbers of photographs and data related to their appeal on web sites such as Flickr, but do not have any information about the daily life of the users who uploaded them. In this paper, we propose a method to close the gap between these two data sources, using the Nokia Mobile Data Challenge (NMDC) dataset[2] to model personal behavior and Flickr images to evaluate appeal.

Mobile sensor data has been used to estimate people's behaviors, specifically moods [6]. Here, we use the geotags in the NMDC dataset to profile people's behavior patterns, which we divide into rare and repeating events (Section 2). We then retrieve geotagged images and their "view" and "favorite" counts from Flickr using GPS and time coordinates of the events. We calculate the ratio of favorite count to view count as the *image appeal rating* of a Flickr image (Section 3). We analyze these two datasets and show that there is a connection between the events in the NMDC dataset and the Flickr image appeal ratings. First, the Flickr images that are taken during the events where the NMDC participant also took photographs have high ratings. Second, the Flickr image appeal ratings are higher for rarer events and lower for repeating events.

Motivated by that, we introduce *event-based features* to estimate the appeal rating of geotagged Flickr images (Section 4). The estimated ratings are used to rank these images. Finally, we compare our ranking with the actual ranking of Flickr images and obtain a Spearman's correlation coefficient of 0.47 (Section 5).

GeoMM'13, October 21, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2509230.2509234.

[1] http://www.flickr.com/
[2] http://research.nokia.com/page/12000

To summarize previous state-of-the-art, research on image appeal prediction can be grouped under visual-based and data-fusion based methods. In visual-based methods, only pixel information is used to estimate the rating. To predict the aesthetic quality of an image, researchers used image composition, contrast, low depth of field [1, 7], and face and object detectors [1]. Due to the semantic gap between the visual features and actual image content, visual-based techniques have a limited performance in rating estimation.

In order to reduce the semantic gap, data-fusion based methods benefit from image tags [2, 5, 8, 10], comments and social interaction between website users [9, 10] on image hosting web sites. The main drawback of these methods is that they require human involvement to tag and comment on images, whereas our method only requires automatically collected geotags such as the ones in the NMDC dataset. Yin et al. [11] estimated the quality of a geotagged input image by using auxiliary photographs that share similar GPS coordinates with the input. Unlike our method, this approach is limited to either popular locations or non-popular locations with scene category assumptions.

2. THE NMDC DATASET

The NMDC dataset includes smartphone usage information and various sensor data collected in Switzerland during approximately two years. In our analyses, we use over 10 million GPS data points (time, location) that belong to 166 NMDC participants. In addition, the participants took a photograph at 17'000 of these GPS data points. In Figure 1, the GPS density map of the participants and the GPS positions of the collected photographs are overlaid on the map of Switzerland. Because the challenge targeted the people around the Lake Geneva region, there is more data for the south-west part of Switzerland.

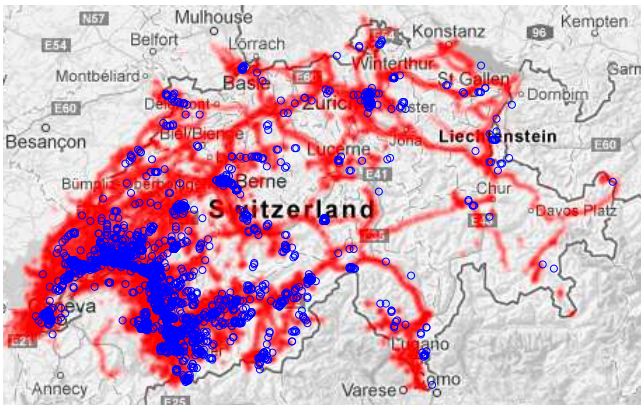


Figure 1: Map of over 10 million GPS coordinates (in red) and 17'000 coordinates of photographs taken by all participants (in blue circles) (Image is taken from Google Maps).

In order to represent the NMDC participants' behavior patterns, we use the GPS data points to extract *time- and location-based events*. We define the following types of events that we will refer to in the rest of the paper:

- **Event:** We define an event as a limited spatial and temporal interval, in which the GPS data points of an NMDC participant do not change significantly[3]. We extract over 90'000 events from 10 million GPS data points. From here on, we will call the i^{th} event of participant n as E_i^n.
- **Rare and Repeating Events:** Events are indicated as "rare", if there exist less than ten events for the **same** participant at close-by GPS coordinates ($\leq$ 1 km). The inverse is true for repeating events. For each participant n, we cluster events according to their locations and quantize their repetitions, which is referred to R_i^n in Table 1.

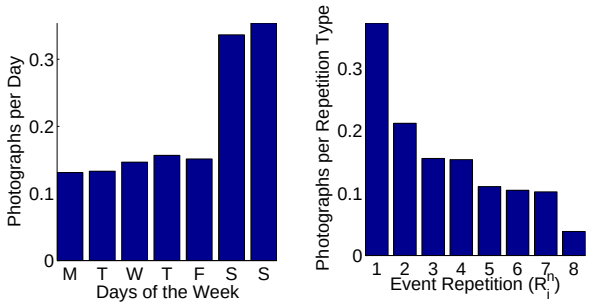
Table 1: R_i^n and corresponding event repetition.

R_i^n	Repetition	Type
1	less than 10	Rare
2	between 10 and 20	Repeating
3	between 20 and 50	
4	between 50 and 100	
5	between 100 and 200	
6	between 200 and 500	
7	between 500 and 1000	
8	between 1000 and 2000	

- **Photographically Interesting and Uninteresting Events:** Events are indicated as "photographically interesting" (PI), if a participant took a photograph during that event. The inverse is true for "photographically uninteresting" (PU) events. We detect approximately 3'700 PI events and 86'300 PU events in the NMDC dataset. The number of photographs collected during an event E_i^n is equal to N_i^n.

A person's behavior pattern can be inferred from the collection of their rare and repeating events. Photographic habits, on the other hand, can be described by PI and PU events.

Each of the 90'000 events has a day, time, center coordinates, repetition (R_i^n), and number of photographs (N_i^n) associated with it. In our analysis, we observe that the days of the week and event repetition are the most influential factors on photograph collection in the NMDC dataset. Figure 2 suggests that the NMDC participants are photographically more active on the weekends. In addition, we can observe that people tend to capture more photographs in the places they visit less frequently (i.e. during rare events).



(a) (b)

Figure 2: Effect of (a) the days of the week and (b) event repetition (R_i^n) on the photograph collection.

3. THE FLICKR DATASET

Due to privacy reasons, the actual photographs that were collected by the NMDC participants were not recorded. Thus, we retrieve over 36'000 geotagged photographs from Flickr. This dataset is formed by downloading all the geotagged photographs within a 100m radius of each event center coordinate and within the NMDC data collection duration of two years. Along with the images, we also get the following Flickr image statistics:

- **View Count:** The number of people who viewed the image (all retrieved images were viewed more than 20 times).
- **Favorite Count:** The number of people who added the image to their favorite images list.

These values can be considered as outputs of a psychophysical test. We will treat the ratio of favorite count to view count as the "image appeal rating" (A_m) of an image I_m.

[3]A significant change corresponds to 10 kilometers for location and 1 hour for time.

In order to investigate the relation between the NMDC and our Flick dataset, we assign every Flickr image to its spatiotemporally closest event[4] and analyze the results. In Figure 3(a), we show the average image appeal ratings of PI and PU events, which are 0.023 and 0.008, respectively. Thus, Flickr users think that the photographs in PI events are 2.71 times more appealing than PU event images.

In addition, we calculate the average image appeal ratings for different event repetitions. As illustrated in Figure 3(b), the images collected during the time and at the locations that the NMDC participants visit less often are more appealing. The relation between the events of the NMDC participants and image appeal ratings (A_m) on Flickr motivates us to build a learning-based image appeal rating predictor.

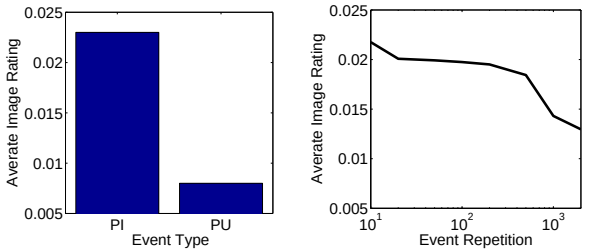


Figure 3: The average image appeal ratings (a) for PI and PU events (b) for different event repetitions.

4. EVENT-BASED FEATURES

The statistical results in Section 3 show that the events of the NMDC participants can contribute to predict the appeal ratings of the images in our Flickr dataset. Thus, we build a framework for estimating the appeal rating of a photograph by using these statistics. We consider that the estimated appeal rating G_m of the m^{th} image I_m in our Flickr dataset as the expected value of Flickr image appeal rating distribution F_m given the image.

$$G_m = \mathbf{E}\big[P(F_m|I_m)\big] = \mathbf{E}\Big[\frac{P(F_m, I_m)}{P(I_m)}\Big] \tag{1}$$

In ideal case, $\mathbf{E}\big[P(F_m|I_m)\big] = A_m$. Every event has a repetition bin R_i^n. We call the set of all events where $R_i^n = k$ as $\mathbf{R_k}$. In Figure 3(b), we see that event repetition is correlated to the image appeal ratings. Thus, we can introduce $\mathbf{R_k}$ as an auxiliary variable to estimate the image appeal rating. Using the chain rule we have

$$\begin{aligned} P(F_m, I_m) &= \sum_{k=1}^{8} P(F_m|I_m, \mathbf{R_k}) P(I_m, \mathbf{R_k}) \\ &= \sum_{k=1}^{8} P(F_m|I_m, \mathbf{R_k}) P(I_m|\mathbf{R_k}) P(\mathbf{R_k}) \end{aligned} \tag{2}$$

The marginal probability of an image $P(I_m)$ is equal for all images. Thus, combining (1) and (2), we have

$$G_m \propto \sum_{k=1}^{8} \mathbf{E}\big[P(F_m|I_m, \mathbf{R_k})\big] P(I_m|\mathbf{R_k}) P(\mathbf{R_k}) \tag{3}$$

The first term in the right-hand side in (3) is effectively estimated during regression training, which is described in Section 5. The second and third terms represent our "event-based features".

[4]10 km spatial distance and 1 hour temporal difference are equivalent in spatiotemporal distance computations

Event-based features express the photographic importance of a location and a time. As no two events can be performed by the same participant at the same position and time, we can rewrite the second term in (3) as follows:

$$P(I_m|\mathbf{R_k}) = \sum_{E_i^n \in \mathbf{R_k}} P(I_m|E_i^n) \tag{4}$$

When we are given a time and location for an image, we compute the probability of having that photograph in an NMDC event. We can relate this probability to an exponential distribution shown in (5). Thus, we effectively favor images that are taken close to an event with a large number of photographs.

$$P(I_m|E_i^n) \propto (N_i^n + \epsilon) \exp\big[-(N_i^n + \epsilon) D(I_m, E_i^n)\big] \tag{5}$$

Here, N_i^n is the number of photographs for event E_i^n, $D(I_m, E_i^n)$ is the spatiotemporal distance between the image I_m and the event E_i^n, and ϵ is equal to 10^{-1} and approximates the exponential distribution to a uniform distribution when $N_i^n = 0$.

In order to compute an event-based feature vector for an image, we go through all the events and find all $P(I_m|E_i^n)$. Every event contributes to a bin according to its repetition, giving 8 features in total (see Table 1). Then, every bin is weighted with the corresponding $P(\mathbf{R_k})$, which can be easily computed by taking the ratio of number of events in set $\mathbf{R_k}$ to the total number of events. An illustration of this feature extraction is given in Figure 4.

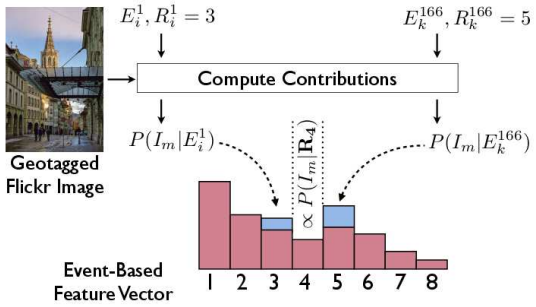


Figure 4: Computation of event-based features of 166 NMDC participants.

In addition to our event-based features, we adopt 24 visual features that perform well in estimating image appeal ratings from the state-of-the-art image appeal rating methods such as [1] and [9]. The explanation of the features are give in Table 2.

Table 2: Visual feature set

Feature	Explanation	Size
Brightness	Mean value of the Y channel in YCbCr space	1
Contrast	Variance of the Y channel in YCbCr space	1
Generalized Contrast	Variance of the color contrast in sRGB space	1
Saturation	Mean and variance of pixelwise color differences	2
Colorfulness	Colorfulness measure in [3]	1
Sharpness	Mean and variance of percentile energy in high frequencies	2
Naturalness	Naturalness measure in [4]	1
Wavelet	3 level wavelet features in HSV space [9]	12
Depth of Field	Center-to-surround wavelet coefficient ratios in HSV space [1]	3

5. EVALUATION AND RESULTS

In our evaluations, we randomly divide the 36'000 Flickr images into equal-sized training and test partitions. We then create 32 regression trees by randomly selecting (with replacement) 80% of the training examples for every tree (i.e. $36000/2 * 0.8 = 14400$ images for each tree). The input of a regression tree is a feature vector, and the ground truth is the image appeal ratings (A_m), which is explained in Section 3. We train regression trees under three setups: visual features only, event-based features only, and both feature vectors. When we test an image to get a final image appeal rating estimation, we pass the extracted features through all trees and average the regression output.

The appeal rating prediction is not directly a curve-fitting problem. Thus, to evaluate the performance of our method, we use rank-correlation metrics instead of standard error metrics such as mean-squared error. For this purpose, we first rank the test images according to their estimated appeal ratings (G_m) and appeal ratings (A_m) and obtain G'_m and A'_m respectively. We then use Spearman's rank correlation test, which measures the correlation coefficient between the estimated and the actual rankings as follows:

$$\rho = 1 - \frac{6 \sum_{m=1}^{K} (G'_m - A'_m)^2}{K^3 - K} \quad (6)$$

Here, K is the total number of test images (i.e. $K = 18'000$). The correlation coefficients for different training setups are given in Table 3.

Table 3: Spearmen's ρ values for individual and combined feature sets averaged over ten experiments.

Method	Spearman's ρ	Extraction Time
Visual	0.3488 ± 0.0050	0.1638 s
Event-Based	$\mathbf{0.4740 \pm 0.0040}$	**0.0043 s**
Visual + Event-Based	$\mathbf{0.4913 \pm 0.0033}$	0.1681 s

As we can see from Table 3, the event-based features are better than visual features in ranking images for our Flickr dataset. It is possible that the pixel-based visual features miss the context information that might implicitly exist in the time and location data. This result is in accordance with the recent research [8, 9, 10] showing that other sources of information, such as text and social interaction, perform better than the visual features in ranking images.

In addition, the event-based features are extracted 40 times faster than the visual features, as they require only a few operations to calculate and they do not depend on the size of the image. This is a very important factor if the number of images that should be ranked reaches millions. In our case, it takes 49 minutes for visual features and 1.3 minutes for event-based features to rank the images for one experiment (regression time $< 10^{-6}s$). However, the best results are obtained by concatenating the two feature sets. The performance boost is statistically significant when we combine the feature vectors. In Figure 6, we present top- and bottom-ranked images in our Flickr dataset using combined feature sets.

In a second experiment, using different thresholds, we divide the test images into two sets with respect to their G_m values. We then estimate the set that a test image belongs to and calculate precision-recall curves shown in Figure 5. Here, the curves for event-based and combined features are similar, which is expected due to the results in Table 3.

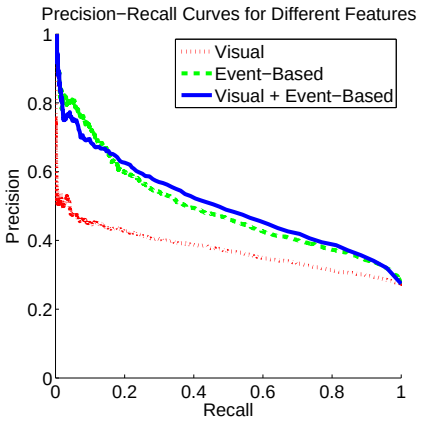


Figure 5: Precision - recall curves for different feature types.

6. CONCLUSIONS & FUTURE WORK

In this paper, we show that photographic habits and behavior patterns retrieved from geotagged mobile data is related to image appeal ratings on photographic web sites. Specifically, we analyze the behavior of the NMDC participants, use time and location to download geotagged Flickr images, and are able to predict their appeal with event-based features. We characterize the NMDC participants' daily lives by defining time- and location-based events. We see that the photographically interesting and rare events are related to the appeal ratings of geotagged Flickr images. We exploit this result by introducing event-based features; they summarize the photographic habits of a collection of people for different times and locations. We show that for our geotagged Flickr photograph dataset, event-based features are more powerful and faster than visual features in ranking image appeal.

The NMDC dataset is limited to locations in Switzerland. Due to the versatility of the method, however, it is possible to extend the algorithm to other locations without any modifications, if similar information is available. In addition, we believe that if the number of participants increases, the accuracy of the method will also increase. By following an individual's daily life and photographic trends, we can retrieve and present more related and appealing images on a personal scale, which will have applications in personal photo albums.

7. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation under grant number 200021_143406 / 1.

8. REFERENCES

[1] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proc. IEEE CVPR*, pages 1657–1664, 2011.

[2] B. Geng, L. Yang, C. Xu, X. Hua, and S. Li. The role of attractiveness in web image search. In *Proc. ACM Multimedia*, pages 63–72, 2011.

[3] D. Hasler and S. Süsstrunk. Measuring colorfulness in natural images. In *Proc. SPIE Human Vision and Electronic Imaging*, volume 5007, pages 87–95, 2003.

[4] K. Huang, Q. Wang, and Z. Wu. Natural color image enhancement and evaluation algorithm based on human visual system. *Computer Vision and Image Understanding*, 103(1):52–63, 2006.

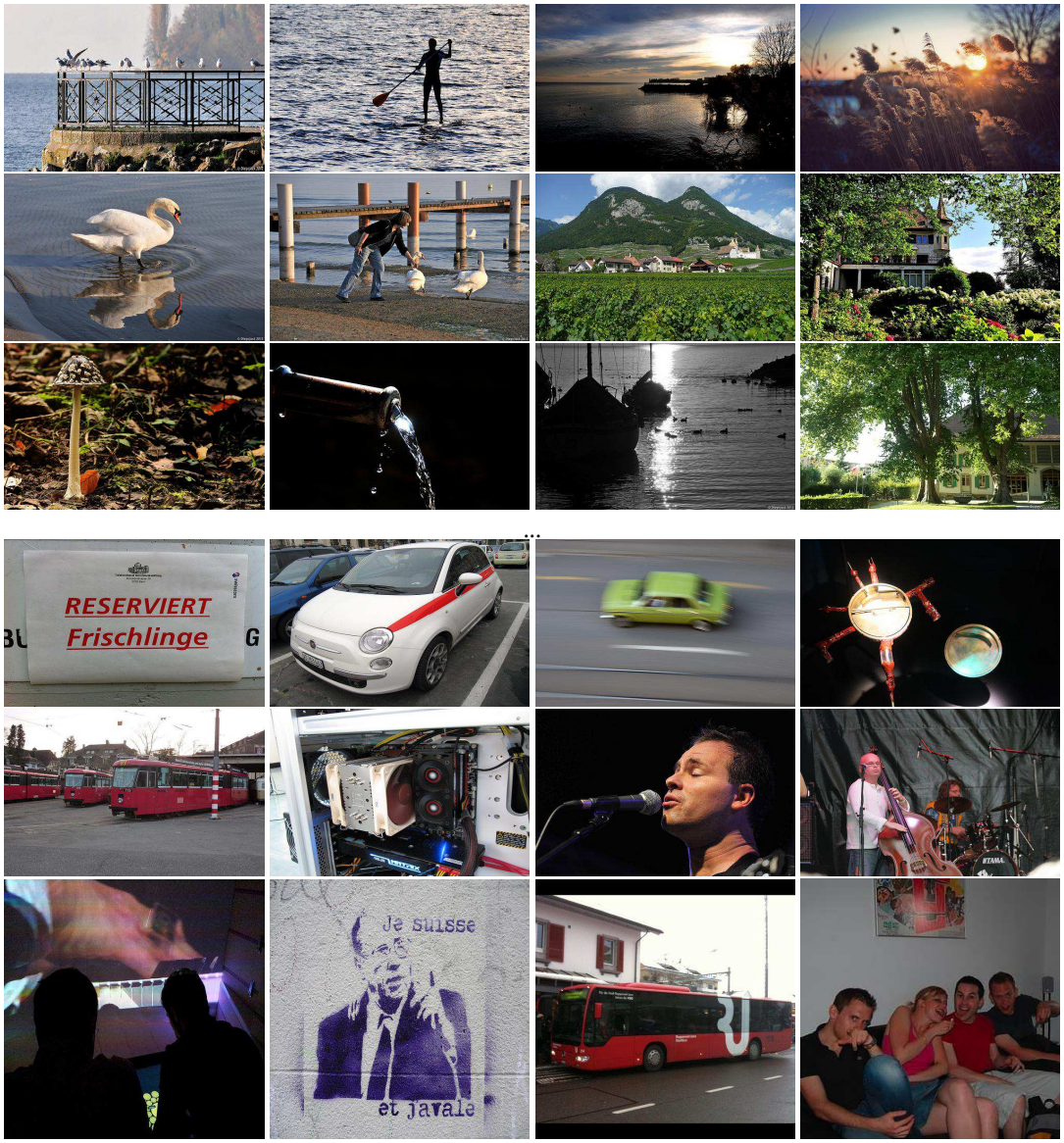


Figure 6: The top- (first three rows) and bottom-ranked (last three rows) images from our Flickr dataset that are sorted by our method using combined features.

[5] J. Jeong, H. Hong, J. Heu, I. Qasim, and D. Lee. Visual summarization of the social image collection using image attractiveness learned from social behaviors. In *Proc. IEEE Multimedia and Expo*, pages 538–543, 2012.

[6] R. LiKamWa, Y. Liu, N. Lane, and L. Zhong. Moodscope building a mood sensor from smartphone usage patterns. In *Proc. ACM Mobile Systems, Applications, and Services*, 2013.

[7] M. Redi and B. Merialdo. "Where is the interestingness?": retrieving appealing videoscenes by learning flickr-based graded judgments. In *Proc. ACM Multimedia*, pages 1363–1364, 2012.

[8] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proc. ACM World Wide Web*, pages 771–780, 2009.

[9] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. In *Proc. ACM World Wide Web*, pages 439–448, 2012.

[10] R. van Zwol, A. Rae, and L. G. Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *Proc. ACM Multimedia*, pages 1015–1018, 2010.

[11] W. Yin, T. Mei, and C. W. Chen. Assessing photo quality with geo-context and crowdsourced photos. In *Proc. Visual Communications and Image Processing*, pages 1–6, 2012.