

Leveraging User Comments for Aesthetic Aware Image Search Reranking

Jose San Pedro^{*}
Telefonica Research
Barcelona, Spain
jspw@tid.es

Tom Yeh
University of Maryland
College Park, Maryland, USA
tomyeh@umd.edu

Nuria Oliver
Telefonica Research
Barcelona, Spain
nuriao@tid.es

ABSTRACT

The increasing number of images available online has created a growing need for efficient ways to search for relevant content. Text-based query search is the most common approach to retrieve images from the Web. In this approach, the similarity between the input query and the metadata of images is used to find relevant information. However, as the amount of available images grows, the number of relevant images also increases, all of them sharing very similar metadata but differing in other visual characteristics. This paper studies the influence of visual aesthetic quality in search results as a complementary attribute to relevance. By considering aesthetics, a new ranking parameter is introduced aimed at improving the quality at the top ranks when large amounts of relevant results exist. Two strategies for aesthetic rating inference are proposed: one based on visual content, another based on the analysis of user comments to detect opinions about the quality of images. The results of a user study with 58 participants show that the comment-based aesthetic predictor outperforms the visual content-based strategy, and reveals that aesthetic-aware rankings are preferred by users searching for photographs on the Web.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems

Keywords

opinion mining, visual aesthetics modeling, image search reranking, user comments, sentiment analysis

1. INTRODUCTION

Billions of digital photographs have been shared in photography-centered online communities, such as Flickr, Facebook or Picassa. The increasing size of photography collections poses a challenge to retrieval algorithms, which need to deal in real-time with these vast sets to find the most relevant

assets. The text query-based approach is the most common for image search. This approach operates on the textual metadata associated with images (e.g. tags, comments, descriptions), reducing the image search task to finding relevant text documents. Text-based image search achieves successful results, especially in online sharing sites where the community devotes significant time to providing quality metadata (e.g. Flickr). However, in many other settings it finds significant shortcomings. For instance, image search engines infer image metadata from their surrounding text in Web pages, which is often noisy. In addition, human provided annotations tend to be sparse and noisy, turning them into an unreliable information source for retrieval [5].

Previous literature has considered image reranking methods aimed at dealing with noisy metadata with the goal of promoting relevant content to the top ranks. A common strategy is to select a group of relevant images from the original result set, and learn content-based models to select similar images [21, 3]. Nevertheless, the increasing size of collections poses an additional challenge: when working at very large scale, the chances of having too many assets similarly relevant to the original query grow. For instance, querying for “dog” would find thousands of relevant images in typical Web image datasets. Increasingly sophisticated ranking and reranking schemes solely based on relevance can deal with the problem only to a certain extent. When too many relevant resources exist in the dataset, additional parameters need to be considered for ranking search results.

In this paper, we focus on the study of an additional aspect to incorporate to the ranking of image search results: visual aesthetic appeal. The pictorial nature of images is responsible for generating intense responses in the human brain, as we are greatly influenced by the perception of our vision system [15]. The aesthetic appeal of images relates to their ability to generate a positive response in human observers. Such a response can be affected by objective and subjective factors, and is able to create important emotional binds between the observer and the image [11].

We focus on the Web image search problem setting and study the influence of visual aesthetic quality in search results. Our hypothesis is that, when searching for images on the Web, users tend to prefer aesthetically pleasant images as long as they remain relevant to the original query. The main contributions of this paper are:

- A method to perform rating inference [14] from user comments about photographs. To this end, we use sentiment analysis tools to extract positive and negative opinions of users, which are then used to train rating inference models, as suggested in [13, 17]. Pre-

^{*}Author was a visiting scholar at The Pennsylvania State University during the realization of this paper.

dicted ratings serve as proxies for aesthetic quality of photographs [16].

- A large-scale user evaluation about the impact of aesthetic-based reranking in the perceived quality of search results. This study is the first to consider aggregated scores combining relevance and aesthetic features to determine the user’s perceived quality of search results.

The paper is organized as follows. We review related literature in Section 2. We describe our rating inference model to predict visual aesthetics by leveraging user’s comments in Section 3. Section 4 presents an additional aesthetic model based on visual features that we use as baseline. Section 5 presents our proposed method to combine relevance and aesthetic features for reranking search results. We evaluate our proposed methods in Section 6. We conclude in Section 7.

2. RELATED WORK

Image search reranking methods have traditionally focused on promoting the ranks of relevant content to improve the results of text-based queries returned by search engines. These methods leverage visual information to deal with the presence of noisy metadata. Classification-based reranking methods use a pseudo-relevance feedback approach [21], where the top and bottom k results are chosen as positive and negative samples in terms of relevance to the current query. These samples serve as training data to build classification and regression models, which are then used to compute a new set of scores to rank the images. Clustering-based reranking methods group images in clusters, and sort them according to their probability of relevance. The largest cluster is commonly assumed to contain the most relevant images, and results are reranked based on the distance to that cluster [3]. In graph-based reranking methods, images are considered nodes in a graph, and edges represent visual connections between them. Edges are assigned weights proportional to their similarity. Reranking can be formalized as a random walk or an energy minimization problem [7].

In this paper, we pursue a different reranking strategy. Our goal is to incorporate alternative aspects into search results ranking that could complement relevance as the only sorting factor. There have been few relevant works in this direction. An interesting approach proposed by Wang *et al.* consists in reranking search results to promote accessibility for colorblind people [20]. Their method effectively demotes images that cannot be correctly perceived by visually impaired people. An alternative ranking approach, and the one we adopt in this paper, is aesthetic-oriented reranking, which aims at promoting the rank of attractive images [11, 10]. Our work follows this same aesthetic-driven approach, but in contrast to previous works we take into account actual text relevance values (in contrast to ordinal rank positions) to combine with aesthetic scores. This is the first study where relevance and aesthetic scores have been jointly used to evaluate the influence of aesthetics in image search.

Aesthetic-oriented reranking requires models to predict the aesthetic value of images. Visual aesthetic modeling has been receiving growing attention, especially from the multimedia and the human-computer interaction research communities, and is normally posed as a rating inference problem [1, 16]. Most works in these fields leverage content-based features from images to infer the quality of aspects related to aesthetics. Composition and framing features have attracted significant attention [12, 22]. Other visual features used for

aesthetic modeling include: perceived depth of field, color contrast and harmony [8], segmentation [22, 11], or shapes [1]. Contextual information has also been leveraged for aesthetic modeling, including tags [16] and social links [19], which significantly outperforms content-based approaches. The analysis of user opinions to create probabilistic rating inference models is a popular research topic (e.g. prediction of movie ratings using IMDb user comments [14, 9]). Their use for predicting photograph ratings, which serve as proxies for aesthetic quality [16], has been previously suggested in [13, 17]. This is the first work in which such an approach has been developed and evaluated.

3. MODELING AESTHETICS FROM USER COMMENTS

The aesthetic value of photographs is a very subjective concept, and therefore poses a big challenge in terms of modeling. However, researchers have agreed on a set of principles that are key in the human perception of aesthetics in relation to photographs [15]. In photography, world scenes are selectively captured, being the task of the photographer to compose the photograph so the main subject of the picture gathers the viewer’s attention. Photography becomes a subtractive effort: the goal is to achieve simplicity by eliminating all potentially distracting elements from the scene. By properly composing and isolating the main subject, good photographs guide their viewers’ eyes, achieving then their main goal: conveying the photographer’s statement.

High quality pictures tend to exploit shallow depths of field captured using wide apertures, which create photographs with very sharp subjects surrounded by out of focus backgrounds (known as *bokeh*). Composition is also fundamental: specific proportion-related rules (e.g. golden ratio, rule of thirds) are known to produce more appealing images. These rules define the optimal position, size and spatial relations for the main subject and the rest of elements in the photograph. Color (e.g. contrast, vividness) as well as coarseness (e.g. sharpness, texture) features have also direct influence over our perception of visual aesthetics.

Most aesthetic inference methods analyze visual content to determine image quality based on these accepted rules. While they achieve relative success, leveraging contextual information (e.g. image tags) outperforms purely visual models [16]. In this paper, we study the use of user comments for photography rating inference [14] as an approach to model aesthetics. This approach enables us to leverage the ability of humans to judge images, possibly a more accurate information source about aesthetic value than visual or other contextual features [13]. In addition, we are able to reveal the commonly agreed set of most relevant features by analyzing their relative frequency of appearance in comments.

3.1 User’s Comments Source

We use a rating inference approach to aesthetic modeling, where user comments are leveraged to predict quality scores for photographs [14]. To this end, we need a dataset of pictures as training data that contains both user comments and ratings. Having both sources of information allows us to model the predictive relationship between aesthetic features extracted from comments and aesthetic scores.

We found DPChallenge¹ to be an online photo sharing collection well suited to our requirements. DPChallenge is a

¹<http://www.dpchallenge.com>








09/17/2011 11:29:25 AM
I love this vanishing point composition with a very well done composition. Exposure balanced and delicious contrasts.
 Photographer found comment helpful.
09/17/2011 09:20:40 AM
Great lines and tones and architecture! I see Waldo.
 Photographer found comment helpful.
09/15/2011 09:31:27 AM
Great photo with wide range of tones and design elements that is nearly hypnotic. Perfectly framed. Excellent.
 Photographer found comment helpful.
09/14/2011 10:49:09 PM
Let's just call it Le Cloitre. Or L'Abbaye
 Photographer found comment helpful.
09/14/2011 08:30:17 PM
Nun does it better than a Waldo (Waldina?) approx 6 pillars down on the right.
 Photographer found comment helpful.
09/14/2011 06:03:17 PM
I can see Waldo.
 Photographer found comment helpful.
09/14/2011 09:08:14 AM
great shadows and placement of figure. keeps the viewer guessing with what sticking out? Good title, well done architecturally.
 Photographer found comment helpful.

Figure 1: Example of a photograph’s comments in DPChallenge. These comments remark that the photograph excels in composition, exposure, contrast, tones and shadow treatment.

website that features weekly digital photography contests about diverse topics, where users submit their best photographs and compete with each other. Challenges are a key component of the site, and constitute an important incentive for user participation. The competitive nature of DPChallenge has attracted a community of mainly professional and serious amateur photographers.

Pictures are primarily uploaded to compete in challenges, in which winners are decided by the votes casted by the community for each participant image. A comprehensive record of votes received (in a 1 to 10 scale), along with average score values, is kept for each photograph. These scores provide a clear indicator of the quality of photographs and have previously been used to predict aesthetic value [8]. In addition to numeric votes, users are allowed to leave feedback in the form of free text comments about the aspects that they like and dislike about the photographs.

We conducted a preliminary study of the characteristics of DPChallenge comments. This study revealed highly valuable qualitative information about technical aspects of the photographs, many of them related to features relevant to their aesthetic quality. An example of comments extracted from DPChallenge is shown in Figure 1. The fact that DPChallenge has both comments and scores gives us an opportunity to learn a comment-based aesthetic model. To this end, we train a regression model using features extracted from comments and voting scores as ground truth, as described in Section 3.3.

3.2 Analysis of Users’ Comments

In this section we describe the analysis tools we use to extract aesthetic quality information from user comments. At the core of our strategy lies a sentiment analysis algorithm, inspired by previous literature on the subject of *Rating Inference* and *Aspect Ranking*. Aspect ranking aims at identifying important aspects of products from consumer reviews using a sentiment classifier [23]. We use the same conceptual

idea to extract the aesthetic features in which photographs stand out by means of mining opinions from user comments, and infer image ratings from them [14, 9, 17].

3.2.1 Background

We extract opinions from user comments using the supervised approach originally presented by Jin *et al.* in [6]. This method was chosen because of: 1) its ability to deal with multiple opinions in the same document, 2) its ability to extract which features are being judged, and 3) its high prediction accuracy. It relies on a comprehensive training pre-stage in which the model learns to classify text tokens as one of the following *entities*:

- **Features:** words that describe specific characteristics of the item being commented. In our problem setting, these would be aspects of photographs, such as color, composition or lighting.
- **Opinions:** ideas and thoughts expressed in a comment about a certain feature of the item. Opinion entities are subdivided into two types: positively and negatively-oriented.
- **Background:** words not directly related to the expression of opinions.

Let us consider the sentence “*Composition is a bit too centered but good lighting*”. The analysis of this sentence would ideally produce the following entity predictions: *Composition (feature)* is a bit *(background)* too centered *(negative)* but *(background)* good *(positive)* lighting *(feature)*.

The problem statement is the following. Given a tokenized sentence, i.e. a sequence of words $W = w_1, \dots, w_n$, the task is to find the sequence of entities, $\hat{T} = t_1, \dots, t_n$, that best represents the sentiment function of each word. This task is performed using lexicalized Hidden Markov Models (HMM), which extend HMMs by integrating linguistic features, such as part-of-speech (POS) tags and lexical patterns. Observable states are represented by duplets (w_i, s_i) , where s_i is defined as the POS of w_i . We define $S = s_1, \dots, s_n$ as the sequence of POS tags for the current phrase W . In this formulation, the problem of finding the best combination of hidden states, \hat{T} , is solved by maximizing the conditional probability $P(T|W, S)$. This probability can be expressed as a function of the complete sequence of markov states. However, in traditional HMMs this expression is simplified by assuming transitional independence: the next state depends only on the current, i.e. $P(t_i|t_1, \dots, t_{i-1}) \approx P(t_i|t_{i-1})$.

In the case of lexicalized HMMs, the last word observed, w_{i-1} , is introduced in the approximation. The rationale behind this is that keeping track of the last word observed could help in the determination of the entity type of the next word. For instance, in the sentence “Tones are too bright”, the adjective *bright* is used to negatively describe the color tones of the picture. But in the sentence “I love how bright the colors are”, *bright* denotes a positive feeling. This example shows how the prediction can be enhanced by considering the precedent word (*too* or *how*). To account for cases not present in the training data, lexicalized parameters are smoothed using their related non-lexicalized probabilities, giving the final formulation:

$$\begin{aligned}
 P'(t_i|w_{i-1}, t_{i-1}) &= \alpha P(t_i|w_{i-1}, t_{i-1}) + (1 - \alpha)P(t_i|t_{i-1}) \\
 P'(w_i|w_{i-1}, s_i, t_i) &= \beta P(w_i|w_{i-1}, s_i, t_i) + \\
 &\quad (1 - \beta)P(w_i|s_i, t_i) \\
 P'(s_i|w_{i-1}, t_i) &= \gamma P(s_i|w_{i-1}, t_i) + (1 - \gamma)P(s_i|t_i)
 \end{aligned}$$

where the interpolation coefficients satisfy $0 \leq \alpha, \beta, \gamma \leq 1$. This smoothing stage endows the algorithm with the ability to predict entity types for word combinations previously unseen, making the technique less sensitive to the comprehensiveness of the training stage. Once these probabilities are estimated, the maximization of the conditional probability $P(T|W, S)$ is obtained using the standard viterbi algorithm. This results in a final sequence \hat{T} of predicted entities for the current phrase.

The algorithm then proceeds to find all the *feature* entities, and assigns them an initial opinion direction using the closest *opinion* entity in the sequence. A simple heuristic approach is used to invert the orientation of the opinion, e.g. from positive to negative, if negation words (e.g. not, don't, didn't) are found within a 5 word range in front of the opinion entity. The final result of the algorithm is a set of duplets (*feature*, $\{-1, +1\}$) summarizing the opinions extracted from the phrase. We denote positively-oriented opinions with the label +1 and negatively-oriented with -1.

3.2.2 Implementation Details

The original method [6] considered the analysis of online consumer reviews. Analyzing user comments poses slight different challenges. One of the most significant differences is the fact that user comments tend to avoid negative opinions, as they might be considered rude by the community. In contrast, consumer reviews give opinions about products, not people or their creations, so negative judgments are more explicit. A preliminary qualitative analysis of the comments in DPChallenge revealed that users are more prone to give advice and constructive feedback (e.g. *I would increase the vibrancy of colors to improve the result*) rather than plain negative feedback (e.g. *The colors are not very vibrant*).

We extended the heuristic approach of dealing with negation words to consider advice-oriented comments. To this end, we add an additional entity, *advice*, to the HMM model. The goal was to leverage the training data to learn common words and expressions used to convey advice, in consonance to how the method learns *opinion* or *feature* words. Typical examples are conditional modal forms, such as *would* or *should*. By following this approach, we took advantage of the characteristics of the lexicalized HMM model to distinguish between the different uses of these common terms.

Two assessors were recruited to tag a set of comments from our collected dataset (see Section 6.1). Both assessors tagged the same set of 1000 comments, and after inspecting the initial set of responses, were instructed to reach a consensus for the comments in which they had disagreed. To remove ambiguity from the training set, we filtered out comments for which consensus could not be reached. The final training set had 935 labeled comments with inter-user agreement $\kappa = 1$. We trained the model using a maximum entropy classifier as our part-of-speech tagger². We followed a grid strategy to optimize the interpolation coefficients, obtaining the following result: $\alpha = 0.9$, $\beta = 0.8$ and $\gamma = 0.8$.

3.3 Learning Aesthetics From Comments

We are aware that the concept of aesthetic appeal is highly subjective and poses a challenge in terms of modeling. However, the amount of user feedback available from DPChallenge results in a large annotated dataset of photographs, with multiple users leaving their feedback for the same photo

in the form of comments and ratings. Hence, we expect that the average of these opinions would yield an aesthetic prediction model that reflects the perception of the community.

The analysis of user comments from the dataset generates for each analyzed picture p_i a set of duplets $S(p_i) = \{(f_k^i, o_k^i)\}$, where $1 \leq k \leq K_i$ and K_i denotes the number of duplets extracted for picture p_i . In this expression, f_k denotes each of the *feature* entities detected in the comments, and o_k its associated *opinion* value, either -1 or +1. Note that sentences where *features* have been detected but *opinions* have not, will not generate any duplets. Note also that having multiple tuples for the same feature, i.e. $f_k^i = f_l^i, k \neq l$, can happen, as different users are likely to comment on the same set of features.

Next, we generate a feature representation suitable for training a supervised machine learning rating prediction model. Given a dataset of N photographs, $D = \{p_i | 1 \leq i \leq N\}$, we determine the complete set of M_C detected comment-based *features*, $F = \{cf_j | 1 \leq j \leq M_C\}$. We define the $N \times M_C$ matrix of comment-based aesthetic representation, $C = c_{ij}$, where $c_{ij} = cs_j^i$, i.e. the aggregated sentiment score for feature j in p_i :

$$cs_j^i = \sum_{k=1}^{M_C} o_k^i, \forall l : f_l^i = cf_j$$

In the previous expression, we take advantage of the convention used to represent negative and positive opinions by -1 and +1 respectively. Each unique feature cf_j is assigned a single comment-based score for each picture p_i , cs_j^i , which is effectively the number of positive comments minus the number negative comments.

In order to predict aesthetic values for new photographs we use a supervised learning paradigm. In particular, we are interested in learning a regression model as our goal is to obtain lists of photos ranked by their appeal. This approach effectively finds the weight of features extracted from comments in the determination of an overall rating for photographs. These ratings serve then as proxies for aesthetic value. To learn the model, we consider a training set $\{(\vec{p}_1, r_1), \dots, (\vec{p}_n, r_n)\}$ of picture feature vectors \vec{p}_i and associated ratings $r_n \in \mathbb{R}$ (obtained directly from the DPChallenge scores). Vectors \vec{p}_i correspond to rows in matrix C . Ground truth scores r_i are extracted from DPChallenge user voting scores, as described in Section 3.1.

We use SV- ϵ regression [18] to build our learning model. SV- ϵ computes a function $f(\vec{x})$ that has a deviation $\leq \epsilon$ from the target relevance values r_i of the training data. For a family of linear functions $\vec{w} \cdot \vec{x} + b$, $\|\vec{w}\|$ is minimized which results in the following optimization problem:

$$\text{minimize } \frac{1}{2} \|\vec{w}\|^2 \quad (1)$$

$$\text{subject to } \begin{cases} r_i - \vec{w}\vec{p}_i - b \leq \epsilon \\ \vec{w}\vec{p}_i + b - r_i \leq \epsilon \end{cases} \quad (2)$$

By means of the learned regression function f , aesthetic values can be predicted for new photographs simply by computing $f(\vec{p})$ for their feature vectors, resulting in a list of photos ranked by aesthetics.

4. VISUAL-BASED AESTHETIC MODELING

For the purpose of the study presented in this paper, we consider two different aesthetic models: the comment-based model, described in Section 3, and a second model based

²Default POS tagger in NLTK (<http://www.nltk.org/>)

on visual features. We aim at using this additional visual-based aesthetic prediction model as a baseline to compare with the results of the comment-based model, both in terms of accuracy and image search reranking user preference.

We create the additional visual-based aesthetic model using state-of-the-art visual features from previous related work on aesthetics modeling. In particular, we use all the 9 features proposed in [16] and 15 additional dimensions from features proposed in [1]. The first 9 features selected include many aspects of image color and coarseness, both aspects of critical importance to perceived attractiveness:

- **Brightness:** determined as the average luminance of the image pixels, $f_1 = \frac{1}{n} \sum_{(x,y)} Y(x,y)$, where n denotes the total number of pixels in the image, and Y the intensity of the luminance channel for pixel (x,y) in the YUV color space.
- **Contrast:** a measure of the relative variation of luminance. Computed using the RMS-contrast expression $f_2 = \frac{1}{n-1} \sum_{(x,y)} (Y(x,y) - f_1)^2$. The generalization of this expression to the sRGB color space, by considering RGB vectors instead of luminance scalars, is used to create f_3 .
- **Saturation:** a measure of color vividness, computed as the average of

$$S(x,y) = \max(R_{xy}, G_{xy}, B_{xy}) - \min(R_{xy}, G_{xy}, B_{xy})$$

for each pixel in the image, where R_{xy} , G_{xy} and B_{xy} denote the color coordinates in the sRGB color space of pixel (x,y) . Two features are extracted for saturation, the average saturation and its variance:

$$f_4 = \frac{1}{n} \sum_{(x,y)} S(x,y), \quad f_5 = \frac{1}{n-1} \sum_{(x,y)} (S(x,y) - f_4)^2$$

- **Colorfulness (f_6):** a measure of color difference against grey, computed using Hasler’s method [2].
- **Sharpness:** a measure of the clarity and level of detail in an image determined as a function of its Laplacian:

$$f_7 = \frac{1}{n} \sum_{x,y} \frac{L(x,y)}{\mu_{xy}}, \quad \text{with } L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$$

$$f_8 = \frac{1}{n-1} \sum_{x,y} \left(\frac{L(x,y)}{\mu_{xy}} - f_7 \right)^2$$

being μ_{xy} the mean luminance around pixel (x,y) .

- **Naturalness (f_9):** a measure of the extent to which colors in the image correspond to colors found in nature. Computed using the method proposed in [4].

The second set of 15 additional dimensions accounts for compositional and subject isolation aspects not covered by the previous features:

- **Wavelet-based texture (f_{10} to f_{22}):** Texture richness is normally considered as a positive aesthetic feature, since repetitive patterns create a richer sense of harmony and perspective depth. Three-level Daubechies wavelets are used to derive 12 visual features in the HSV color space. For each level ($l=1,2,3$) and channel ($c=H,S,V$) we compute the following nine features:

$$f_{l,c} = \frac{1}{S_l} \left\{ \sum_{b \in \{LH,HL,HH\}} \sum_{(x,y) \in b} w_{l,c}^b(x,y) \right\}$$

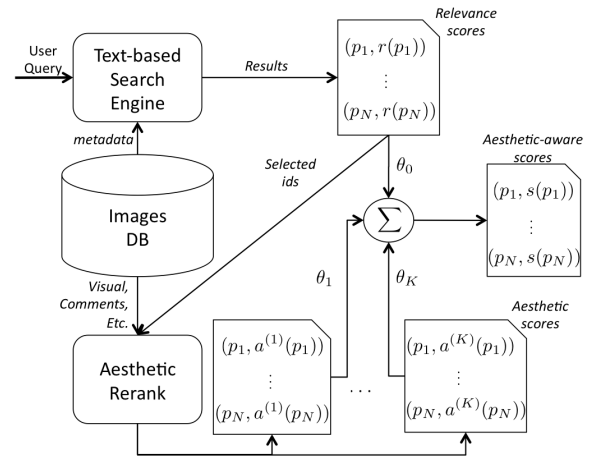


Figure 2: Reranking strategy. Relevance scores are produced from image metadata. Images selected by relevance are used to create K different aesthetic scores derived from different predictors. All scores are then combined to generate the final ranking.

where S_l denotes the size of the level l , b denotes the wavelet higher frequency subbands (LH,HL,HH), and $w_{l,c}^b$ denotes the wavelet transformed values for the given level l , subband b and channel c . Average values for each channel HSV, at all levels l , are used to compute 3 additional features.

- **Depth of Field (f_{23} to f_{25}):** Shallow depths of field are used to separate the main subject from the background. Images are split into 16 equal rectangular blocks, M_1 to M_{16} , numbered from left-to-right, top-to-bottom. The DOF feature is then defined as:

$$f_{DOF} = \frac{\sum_{(x,y) \in M_6 \cup M_7 \cup M_{10} \cup M_{11}} w_3(x,y)}{\sum_{i=0}^{16} \sum_{(x,y) \in M_i} w_3(x,y)}$$

where w_3 denotes the 3-level Daubechies wavelet for the higher frequency subbands (LH,HL and HH). This feature detects objects in focus centered in the frame against an out of focus background. It is computed for each of the three channels in the HSV color space.

Using these 25 features, we build a $N \times 25$ matrix V for denoting the visual-based feature representation for aesthetic modeling, in the same spirit of matrix C (Section 3.3).

5. RERANKING FOR AESTHETICS

This paper studies the impact of aesthetic characteristics of images on the perceived quality of search results by users. To this end, we combine relevance scores obtained by relevance-oriented rank methods with aesthetic quality scores predicted for photographs. We call this combination of relevance and aesthetic scores for ranking *aesthetic-aware reranking*. Intuitively, relevance and aesthetic quality are orthogonal dimensions and therefore convey complementary information about documents being retrieved. In the simplest case scenario, we can think of aesthetic quality as a way to break relevance score ties to enhance results. In this section, we introduce and describe the main components of the reranking strategy adopted, which is illustrated in figure 2.

5.1 Generation of Relevance Scores

Our proposed method takes a list of search results ranked by relevance and rerank them by factoring in aesthetic properties. Relevance scores are generated using a text-based retrieval approach to match the query terms with metadata from the images (e.g. tags, title, description). The most common approach to the computation of relevance scores is based on term frequency-inverse document frequency (tf-idf). Words in queries and documents are often subject to a series of normalizing pre-processing such as stemming, part-of-speech tagging, and stop-word removal. Given a set of query terms, a document is considered more relevant with respect to these query terms if these terms appear more frequently in this document (tf) and fewer other documents also contains these terms (idf).

Note that the proposed approach does not depend on the nature of the original query. Our approach is also valid for query-by-example and query-by-sketch image search paradigms, as it leverages final relevance scores. For this reason, the use of relevance-oriented visual reranking methods prior to the aesthetic reranking stage is also allowed. Therefore, we can effectively combine different reranking strategies focusing on different quality aspects of the search results.

5.2 Aesthetic Value Prediction

The text-based search stage generates a list of retrieved images along with their relevance score for the given query. We predict the visual aesthetic score for each element of the set of retrieved images. In our scenario, we create two different aesthetic scores: one based on the comment-based model, and a second based on the visual-based model. In the final stage of the search process, we combine the original text-based relevance scores (Section 5.1) with the aesthetic values predicted for each image in the result set (Sections 3 and 4). To this end, we use a linear combination model following the expression:

$$s(p_i) = \theta_0 r(p_i) + \sum_{j=1}^K \theta_j a^{(j)}(p_i) \quad (3)$$

where $s(p_i)$ denotes the final combined score for image p_i , $r(p_i)$ its relevance score obtained by the text search engine, and $a^{(j)}(p_i)$ denotes the aesthetic value predicted by the j -th regression model. All scores are assumed to be normalized to take real values in the range $[0, 1]$. This reranking strategy scales well for large-scale collections as aesthetic scores can be updated offline and are not subject to change frequently.

Equation 3 can be tuned to study the independent effects of each aesthetic model, as well as the different possible interactions between them. Section 6.3 provides a large scale user evaluation of the impact of aesthetic-aware reranking in terms of perceived quality of search results. Our user study focuses only on the independent effect of each of the two aesthetic prediction models presented. Therefore, we only combine two rank scores at a time: 1) the relevance-based and 2) either one of the predicted aesthetic scores. We opted for weighting equally relevance and aesthetic scores for the purpose of establishing the potential gain in terms of user satisfaction. Hence, we used $\theta_0 = \theta_1 = 0.5$.

The study of optimization strategies for parameters θ_i lies out of the scope of this paper. These reranking parameters can be used for personalization of search results, where θ_i are dynamically adapted based on historic user click logs. Beyond personalization, we may find additional optimiza-

tion strategies, including adapting θ_i values to the type or content of queries, as suggested in Section 6.3.

6. EXPERIMENTS

This paper contributes (1) a method to predict the aesthetic value of photographs from user comments and (2) evidence that aesthetic-based reranking influences the perceived quality of search results. We conducted experiments to validate and support these two contributions in Sections 6.2 and 6.3 respectively.

6.1 Collected Dataset

We crawled the DPChallenge website and used this collection as our dataset for evaluation. We obtained all available images and metadata from the site, which at the time of the crawl counted with 627,908 photographs. We collected the following information:

- Descriptive metadata: including title, description, and assigned galleries. This information was used to perform the text-based search stage that generates the initial relevance values. We implemented such text-based search engine using the Java Lucene library. We used the standard Porter Stemmer to remove morphological and inflectional endings of all words in the documents and indexed the resulting documents.
- User feedback: including voting scores, which served as ground truth to train the aesthetic models, and user's comments, which we used to build the comment-based representation of photographs. Detailed in Section 3.1.
- Visual information: image files, which we used to build the visual-based feature representation of photographs.

An inspection of the dataset revealed that 64% of the photographs had one or more comments, with a median value of 6. Ratings are less frequent, only present in 41% of the collection. This is caused by DPChallenge limiting votes to photographs that take part in challenges. We only use ratings as ground truth to train inference models, so we are not constrained to this subset for the reranking study.

6.2 Accuracy of Aesthetic Prediction Models

We propose a method to learn a predictive model of visual aesthetics from user comments, which we intend to use for reranking image search results. We pose this as a *rating inference* problem, where predicted photographs ratings serve as proxies for aesthetic quality. We conducted a test to measure its accuracy, and compared it to the purely visual-based model. To this end, we subsampled the DPChallenge dataset uniformly at random and obtained a subset with the following characteristics:

- Training set size: a total of 70,000 photographs, approximately 10% of the full dataset. Photographs with no votes were ignored. In contrast to the complete collection, a higher median value of 9 comments per photograph were available for this subset (mean of 11.45). Performance metrics were obtained using 50,000 randomly sampled items as the training set, and the remaining as the test set.
- Ground truth: scores were extracted from community provided votes, as described in Section 3.
- Comment features: we restricted comment-derived features to those appearing in at least 2% of the photographs of the training set. By considering only the

Table 1: Most popular features extracted by the aesthetic predictor from user comments of the DPChallenge subset. These aspects are the most frequently referenced when users comment on photographs.

color	composition	sharpness	framing	cropping	exposure	dof	tone	lighting	contrast
focus	reflection	processing	shadows	saturation	texture	edges	detail	perspective	angle
subjects	portrait	highlights	model	people	trees	hand	eyes	place	pose
macro	message	execution	idea	abstract	photograph	sense	effort	camera	day
thanks	interpretation	comments	critique	things	stuff	club	part	rest	

most commonly commented features, we reduce both the sparseness and the complexity of the training and the prediction tasks. Table 1 shows the final list of $M_C = 49$ features used for the comment-based aesthetic prediction. A preliminary analysis of these features reveals many aspects related to technical quality of images (e.g. composition, saturation), presence of interesting elements (e.g. portrait, eyes) as well as other higher level aspects (e.g. idea, message).

In terms of feature-representation, we used the following three schemes to build aesthetic prediction models:

- Visual only, **V**: We use the 25 dimensions of matrix V as defined in Section 4
- Comments only, **C**: We use the 49 dimensions of matrix C as defined in Section 3.3.
- Visual + Comments, **VC**: We combine matrices V and C into a single joint representation of visual and comment features. Matrix VC is a $N \times 74$ matrix, $N = 70,000$, with elements:

$$vc_{ij} = \begin{cases} v_{ij} & 1 \leq j \leq 25, \\ c_{i(j-25)} & 26 \leq j \leq 74 \end{cases}$$

We used the R-Squared metric, as well as Spearman’s ρ and Kendall’s τ correlation, as quality metrics. R-Squared is a widely used metric to test models’ goodness of fit based on the aggregated prediction error. Spearman’s ρ and Kendall’s τ are metrics of rank correlation. They provide a measure of the prediction power of models by looking at rank differences when sorting elements by the observed and the predicted values. These latter metrics are more suitable for our problem setting, as we aim at establishing an order between pictures based on their aesthetic value.

Table 2 shows the accuracy values obtained. All correlation values were significant (p-value=0.001). Visual features obtained predictions with relatively low correlation values that are in consonance with previous works in the aesthetic prediction field [16]. Using our method of prediction based on the analysis of user comments, we obtained consistently higher scores for all accuracy metrics. In contrast to visual-based aesthetic modeling, the comment-based representation conveys much higher level information about the quality of pictures. As shown in Figure 1, the model handles information about high level aspects, including the photograph’s *message*, or the subject’s *eyes* and *pose*.

The combination of both sources of information, visual and comments, led to marginal improvements over the comment-based strategy, also in consonance with results found by previous works [16]. This result supports the viability of our approach to use automatic analysis of user comments to predict accurately ratings of photographs. Furthermore, although this approach requires the presence of user comments, we have shown that a dataset featuring a median of 9 comments per photograph achieves high prediction accuracy. We believe that, given the increasing trend of user

Table 2: Accuracy of aesthetic prediction models for visual-based (V), comment-based (C) and combined (VC) feature representations. VC obtains the higher accuracy scores (boldfaced) for all metrics.

	V	C	VC
R-Squared	0.0988	0.3726	0.39889
Spearman’s ρ	0.3133	0.5839	0.6107
Kendall’s τ	0.2125	0.3726	0.4352

participation on the Web, such an amount of comments is likely to be available for large collections of photographs.

6.3 Aesthetic-aware Reranking: User Study

6.3.1 Participants

The hypothesis that drives our work is that, *when searching for images on the Web, users tend to prefer aesthetically pleasant images as long as they remain relevant to the original query*. We conducted a user study to test this hypothesis with 58 participants (32 female) whose ages ranged from 22 to 71 years old (mean 51.15 years). Participants were asked about their knowledge of photography: 25 reported to have “passing knowledge”, 28 to be “knowledgeable” and 4 stated to be “experts”. Participants were recruited using mailing lists and social networks to disseminate the study information. They held a variety of occupations, including researchers, engineers, students and sociologists. The experiment was implemented as a website accessible to participants online. We used the results of participants who completed the session, which took 21 minutes in average.

6.3.2 Methodology

We compiled a set of 25 keywords to study reranking results in variety of cases. These keywords were: people, sky, tree, portrait, flower, building, sunset, car, beach, bird, road, dog, cat, fish, baby, school, horse, food, game, apple, animal, boy, star, heart, and weather. These 25 keywords were selected from a larger set that combined queries used for evaluation in [20] and [24]. The combination of these two collections of queries contained 92 different keywords. We discarded those returning less than 100 elements in our dataset, and manually clustered the remaining in 9 categories according to their topic (animals, plants, landscape, human, sports, travel, food, architecture, miscellaneous). We sorted the keywords in each category by the number of results retrieved in our dataset. We kept the top half of keywords in each category, aiming at having a diverse set of topics. Finally, we used our Lucene-based search engine to find the list of relevant results from our DPChallenge collection for each of these 25 keywords, and kept their top 100 results.

To avoid fatigue, participants were asked to provide their judgments for 15 image search queries randomly selected from the set of 25. For each of the 15 queries, participants

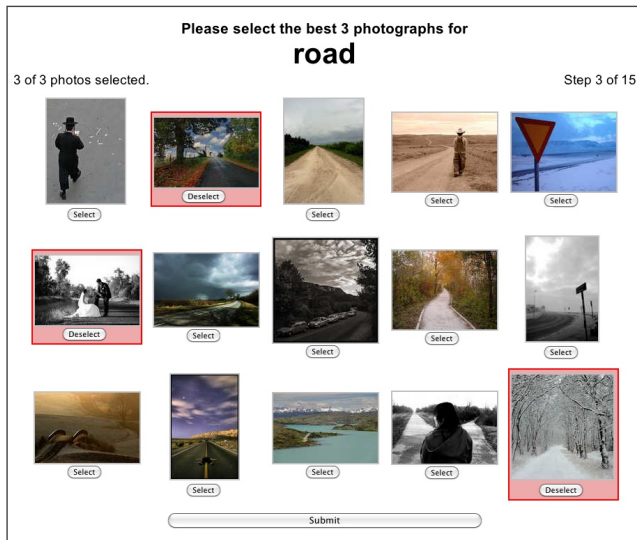


Figure 3: Web interface used by participants in the user study.

were shown what we called the *evaluation set*. The evaluation set consisted of 15 result images from which participants were asked to select the best 3, as illustrated in Figure 3. The 15 images shown for each query had been generated by 3 different ranking strategies without their knowledge: the original text relevance-based rank, an aesthetic-aware reranking based on visual features, and finally an aesthetic-aware reranking based on user comments. We selected these 3 ranking strategies to compare the independent effect in user preference of the two considered aesthetic models to rank search results, along with the original relevance rank.

The aesthetic-aware rankings were generated from the top 100 results retrieved by relevance for each query. As described in Section 5.2, relevance and aesthetic scores were combined, for each strategy, using the proposed lineal combination model with parameters $\theta_0 = \theta_1 = 0.5$. Figure 4 shows the images selected by each ranking strategy to create the evaluation set for two different queries. The 15 images for each evaluation sets were chosen by taking the top 5 images from each ranking strategy. In case of collisions, i.e. when the same image was within the top 5 results of more than one strategy, we selected additional images (rank 6 and below) from all ranking strategies following a random order.

We built a Web interface to conduct this study, which is depicted in Figure 3. Participants could clearly see the search keyword at the top of the screen, and a grid containing the thumbnails of the 15 images just below. Users could click on any image to see a full-size version, and could use the buttons below the thumbnails to select/deselect their chosen ones. To prevent ordering bias, each evaluation set was randomly shuffled.

In order to evaluate the performance of each ranking strategy, we used the metric proposed in [11]. This is computed as the average of two measures:

- Winner Ranking: Quantifies the number of times that selected photos came from each of the three ranking strategies. For each ranking strategy, $i \in \{1, 2, 3\}$, we

Table 3: Results of the ranking preference user study. Each row provides the overall performance metric value, cm_i , for each of the three ranking strategies. The highest value for each query is bold-faced. Rankings tied with the boldfaced winning strategy for each query (difference is not significant at significance level $\alpha = 0.05$) have been shaded.

Query	Ranks		
	Relevance	Visual	Comments
animal	0.2488	0.2825	0.7627
apple	0.4156	0.5471	0.5199
baby	0.5388	0.4181	0.6791
beach	0.5534	0.4440	0.4580
bird	0.4802	0.5728	0.5044
boy	0.4997	0.6659	0.5104
building	0.4791	0.5561	0.5349
car	0.6129	0.4718	0.4307
cat	0.4367	0.5350	0.5720
dog	0.5034	0.4788	0.5266
fish	0.4392	0.5573	0.5750
flower	0.4761	0.4873	0.5724
food	0.5180	0.4998	0.5325
game	0.4596	0.5500	0.6357
heart	0.4348	0.4875	0.7051
horse	0.6281	0.3837	0.5586
people	0.5583	0.3929	0.5580
portrait	0.3995	0.4964	0.5412
road	0.4455	0.5160	0.5745
school	0.6395	0.5920	0.5285
sky	0.3639	0.3722	0.5738
star	0.5540	0.4740	0.5458
sunset	0.4049	0.6622	0.5044
tree	0.4377	0.6338	0.4763
weather	0.5669	0.4137	0.4758
Aggregated	0.4830	0.5010	0.5530
Wins	7	6	12
Tied Wins	11	16	20

compute this score using

$$tm_i = \sum_{j \in \{p_i\}} \frac{I_i(j)}{3} \times \frac{1}{\sum_{k=1}^3 I_k(j)}$$

where $\{p_i\}$ is the set of pictures selected for rank strategy i , and $I_i(j)$ is 1 when image j has been selected by both the user and the i -th rank strategy, and 0 otherwise. The second factor of this equation accounts for collisions between different ranking strategies. The expression $\sum_{i=1}^3 tm_i = 1$ should always hold true.

- Ranking Performance: Quantifies how well each strategy ranked images selected by users. In this case, we compute the position of the 3 chosen pictures within each ranking to compute the score:

$$rm_i = \frac{1}{3} \left(\sum_{j=1}^3 \frac{S - Pos_i(p_j)}{S - j} \right)$$

where $Pos_i(p_j)$ is the position in which the ranking strategy i ranked the user selected picture p_j such that $Pos_i(p_1) < Pos_i(p_2) < Pos_i(p_3)$, and S is the maximum rank considered. We chose $S = 40$ [11].

We compute the overall performance of rank strategy i as

$$cm_i = \frac{tm_i + rm_i}{2}$$

6.3.3 Results

Table 3 shows the performance obtained for each query and proposed ranking strategy in the user study. In the aggregated comparison for the 25 queries, our proposed comment-based aesthetic reranking strategy obtained a higher overall performance score (0.5530) than the relevance and visual-based strategies. We ran an ANOVA and Tukey’s significant difference test (HSD), which revealed that the difference between the comments and the visual and relevance rankings was statistically significant (p-value=0.001).

This significant difference in ranking performance supports the hypothesis that aesthetically pleasant photographs, as selected by the comment-based aesthetic reranking, are preferred by users. Hence, aesthetic aware rankings, which promote the rank aesthetic images, are likely to increase user satisfaction with search results over the original relevance-based ranking. Moreover, the comment-based strategy also performs significantly better than the baseline aesthetic-aware rank based on visual features. This result is in consonance with the better accuracy performance of comment features discussed in Section 6.2. The difference between the visual and relevance rankings resulted in a p-value of 0.0819, not statistically significant at $\alpha = 0.05$.

The analysis of performance for individual queries also revealed a clear predominance of our comment-based approach, which obtained the overall highest score in almost 50% of the cases. Furthermore, its performance was not significantly different from the best strategy in 80% of the queries (at significance level $\alpha = 0.5$). We also found that users felt more inclined towards aesthetic-aware rankings which combine both relevance and aesthetic scores. In 24 of the queries, an aesthetic-aware reranking was preferred (or not significantly different from the preferred choice), with “car” being the only exception.

We observed a noticeable inter-query variation of ranking strategy preference, which suggests that further optimization strategies should be pursued to adapt the weights of each ranking score to the type of query. Image search engines could use the performance metric cm_i to tune the model parameters θ_i for combining image scores.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown that community feedback found in Web-based social sharing systems can be used to improve the ranking of image search results. More specifically, we have leveraged user comments about photographs to create a comment-based feature representation of images conveying the opinion, positive or negative, of users about the images. We have used these features for building regression models aimed at predicting the aesthetic quality of images, using ratings provided by users of the community as ground truth. Finally, we have studied how to combine relevance and aesthetic scores to rerank image search results. Our experiments have shown that context-based representations outperform visual-based in terms of prediction accuracy. We also conducted a user study to determine user satisfaction with aesthetic-aware reranking of search results, which revealed a consistent preference of results reranked by the combination of aesthetic and relevance scores.

We plan to extend this work to consider additional contextual information to improve aesthetic prediction accuracy. One of the most interesting lines of work in this regard is the analysis of social features in the dataset, aiming at weighting comments by the reputation of their authors. Additional contextual cues could be used to extend the feature representation, such as tags or category/topic of photographs. We also plan to study the scalability of the solution as well as the viability of this approach to model aesthetics from user opinions in non-specialized communities, where comments could be less technical and not as well written.

We also want to extend this approach to non-visual domains. For instance, similar quality metrics for text documents could be derived from correlations with attractive topics, sentence structure analysis or vocabulary distributions. In addition, we plan to conduct a large scale qualitative study to determine the reasons behind the preference for different ranking strategies depending on the query.

8. ACKNOWLEDGMENTS

This research was part of the project MIESON. The project MIESON (grant agreement n. 254370) is supported by the European Union under a Marie Curie International Outgoing Fellowship for Career Development.

9. REFERENCES

- [1] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV’06*, volume 3953 of *L.N. in Computer Science*, pages 288–301. Springer.
- [2] S. Hasler and S. Susstrunk. Measuring colorfulness in real images. volume 5007, pages 87–95, 2003.
- [3] W. H. Hsu, L. S. Kennedy, and S. F. Chang. Video search reranking via information bottleneck principle. In *ACM Multimedia ’06*, pages 35–44, NY, USA, 2006.
- [4] K. Q. Huang, Q. Wang, and Z. Y. Wu. Natural color image enhancement and evaluation algorithm based on human visual system. *Comput. Vis. Image Underst.*, 103(1):52–63, 2006.
- [5] V. Jain and M. Varma. Learning to re-rank: query-dependent image re-ranking using click data. In *Proc. ACM Conf. on World wide web*, WWW ’11, pages 277–286, NY, USA, 2011. ACM.
- [6] W. Jin, H. H. Ho, and R. K. Srihari. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proc. ACM SIGKDD*, KDD ’09, pages 1195–1204, NY, USA, 2009. ACM.
- [7] Y. Jing and S. Baluja. Pagerank for product image search. In *Proc. ACM Conf. on World Wide Web*, WWW ’08, pages 307–316, NY, USA, 2008. ACM.
- [8] Y. Ke, X. Tang, and F. Jing. The Design of High-Level Features for Photo Quality Assessment. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:419–426, June 2006.
- [9] C. W. K. Leung, S. C. F. Chan, F. L. Chung, and G. Ngai. A probabilistic rating inference framework for mining user preferences from reviews. *World Wide Web*, 14(2):187–215, Mar. 2011.
- [10] Y. Luo and X. Tang. Photo and Video Quality Evaluation: Focusing on the Subject. In *ECCV ’08*, pages 386–399, Berlin, Heidelberg, 2008. Springer-Verlag.

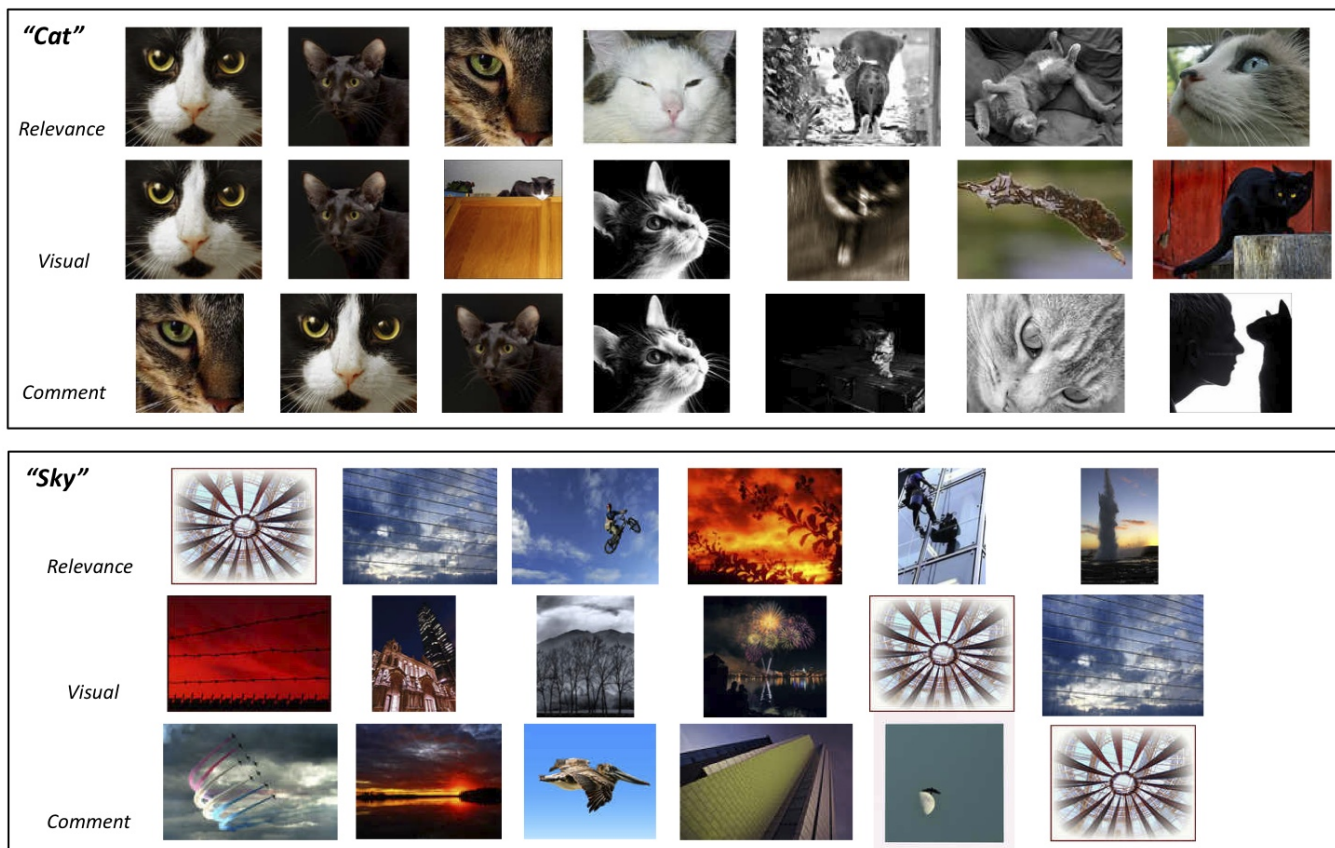


Figure 4: Selection of images for the evaluation set of the queries “cat” and “sky”. Images are sorted from left to right in descending rank score, for each of the 3 ranks considered.

- [11] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver. The role of tags and image aesthetics in social image search. In *WSM '09*, pages 65–72, NY, USA, 2009. ACM.
- [12] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *IEEE ICIP 2010*, pages 3185–3188, 2010.
- [13] R. Orendovici and J. Z. Wang. Training data collection system for a learning-based photographic aesthetic quality inference engine. In *ACM Multimedia '10*, pages 1575–1578, NY, USA, 2010.
- [14] B. Pang and L. Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [15] G. Peters. Aesthetic Primitives of Images for Visualization. In *IEEE Int. Conf. Information Visualization, 2007*, pages 316–325, July 2007.
- [16] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proc. ACM conf. on World wide web, WWW '09*, pages 771–780, NY, USA, 2009.
- [17] N. Sawant, J. Li, and J. Z. Wang. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools Appl.*, 51(1):213–246, 2011.
- [18] A. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug. 2004.
- [19] R. van Zwol, A. Rae, and L. G. Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *ACM Multimedia'10*, pages 1015–1018, NY, USA, 2010.
- [20] M. Wang, B. Liu, and X. S. Hua. Accessible image search. In *ACM Multimedia'09*, pages 291–300, NY, USA, 2009.
- [21] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *ACM Multimedia'10*, pages 183–192, NY, USA, 2010.
- [22] C. H. Yeh, Y. C. Ho, B. A. Barsky, and M. Ouhyoung. Personalized photograph ranking and selection system. In *ACM Multimedia'10*, pages 211–220, NY, USA, 2010.
- [23] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua. Aspect ranking : Identifying important product aspects from online consumer reviews. *Computational Linguistics*, pages 1496–1505, 2011.
- [24] Z. J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T. S. Chua, and X. S. Hua. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Trans. Multimedia Comput. Commun. Appl.*, 6, Aug. 2010.