Ranking and Classifying Attractiveness of Photos in Folksonomies

Jose San Pedro
University of Sheffield
211 Portobello Street
Sheffield S1 4DP, UK
jsanpedro@mac.com

Stefan Siersdorfer
L3S Research Center. University of Hannover.
Appelstr. 9a
30167 Hannover, Germany
siersdorfer@L3S.de

ABSTRACT

Web 2.0 applications like Flickr, YouTube, or Del.icio.us are increasingly popular online communities for creating, editing and sharing content. The growing size of these folksonomies poses new challenges in terms of search and data mining. In this paper we introduce a novel methodology for automatically ranking and classifying photos according to their attractiveness for folksonomy members. To this end, we exploit image features known for having significant effects on the visual quality perceived by humans (e.g. sharpness and colorfulness) as well as textual meta data, in what is a multi-modal approach. Using feedback and annotations available in the Web 2.0 photo sharing system Flickr, we assign relevance values to the photos and train classification and regression models based on these relevance assignments. With the resulting machine learning models we categorize and rank photos according to their attractiveness. Applications include enhanced ranking functions for search and recommender methods for attractive content. Large scale experiments on a collection of Flickr photos demonstrate the viability of our approach.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous; H.3.5 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL—*On-line Information Services*

General Terms

Algorithms

Keywords

image analysis, attractiveness features, photo appeal, web 2.0, classification, ranking, folksonomy feedback

1. INTRODUCTION

Popularity and data volume of modern Web 2.0 content sharing applications originate in their ease of operation for even unexperienced users, suitable mechanisms for supporting collaboration, and attractiveness of shared annotated material (images in Flickr, videos in YouTube, bookmarks in del.icio.us, etc.).

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

The rapid increase in size of online communities and the availability of large amounts of shared data make discovering relevant content and finding related users a difficult task. For instance, thousands of new photos are uploaded to Flickr every minute making effective automatic content filtering techniques a necessity.

Flickr photos are accompanied by a variety of meta data such as tags, number of views, user comments, upload date, etc. The Flickr search interface exploits the explicit and implicit ratings in the meta data to infer rankings. For instance, the number of views is an indicator for the popularity of a photo, the upload date and the date a photo was taken at is an indicator for the recency of the content, and adding a photo to one's favorite list is probably the most direct positive relevance assignment in Flickr, and is an explicit expression of interest in the photo. However, for recently uploaded photos community feedback in any form might not yet be available. Furthermore, many photos are just sparsely annotated which might prevent text-based search and mining methods from retrieving this potentially attractive content.

Visual attractiveness is a highly subjective concept which has received extensive interest from the research community. Semantic aspects are not critical, as the presence of certain concepts in a picture does not necessarily correlate with its appeal for viewers. The artistic component has a major role in the perception of the aesthetics of images, and low-level features can provide a better insight on this aspect of photos. Metrics such as sharpness, an overall value of the granularity of the image, or colorfulness, which measures the diversity of spectrum contained in the image, have been shown to provide high correlation with the human perception of attractiveness ([24]).

In this paper we focus on a methodology for automatically classifying and ranking photos according to their attractiveness. We exploit the vast amount of social feedback available in Web 2.0 applications, more specifically in Flickr, to obtain a training set of photos considered as more or less attractive by the community. This allows us to build classification and regression models based on multi-modal visual and textual features, and to apply them to identify new attractive content. In a wider system context, such techniques can be useful to enhance ranking functions for photo search, and, more generally, to complement mining and retrieval methods based on text, other meta data and social dimensions.

The rest of this paper is organized as follows: In Section 2 we discuss related work on image features, visual attractiveness, folksonomy mining, and machine learning. Section 3 provides an overview of image attributes commonly asso-

ciated to the human perception of visual quality, and reviews methods to compute their values. We provide a short overview of classification and regression techniques in Section 4, and explain how we can apply these techniques in the context of photo attractiveness detection. In Section 5 we provide the results of the evaluation of our automatic attractiveness detection methods for classification and ranking of photos in Flickr. We conclude and show directions of our future work in Section 6.

2. RELATED WORK

The determination of image quality metrics has received significant research interest under very diverse lines of work. The analysis of human perception of color signals is the basis for an important number of image transformation techniques, as it provides a mechanism to assess visual quality at a perceptual level, i.e. as it is perceived by a human observer. This is a common requirement for the evaluation of image compression algorithms [32], but also has applications in image enhancement techniques [12] and unsupervised calibration systems for image capturing gear (e.g. auto-focus systems [30]). Classic metrics such as PSNR and MSE [33] model quality degradation as a measure of the difference between a baseline image and a variation. They perform poorly as objective quality metrics, as they neglect the perceptual impairments associated to the absolute changes in signal values. In the scope of this paper we are interested in quantitative metrics of perceived image quality rather than visual fidelity.

Savakis *et al.* present a subjective evaluation of the significance of different visual aspects for the determination of the overall appeal of natural images, in this case consumer photographs [24]. Their results show that, while the main factors are related to the presence of determinate concepts (e.g. people) and artistic value (e.g. composition), some specific objective measures of visual features provide significant correlation to human judgements. These results are supported by the work of Winkler [31] and Wee *et al.* [30], who propose ways to quantify sharpness and colorfulness of images and conduct extensive subjective experiments showing the properties of these features as effective indicators of image appeal. Additional metrics such as exposure [25], contrast [33, 22], or texture features [20] have also been used with varying levels of success to provide metrics of image appeal. These metrics have been exploited for image retrieval and management applications, e.g. detection and removal of undesirable images from photo collections [25].

However, few works have focused on providing accurate statistical models combining multiple features to predict the attractiveness of images. Kalenova *et al.* [16] propose an unsupervised model for spectral image quality characterization, and considers a very restricted set of only 5 images for its evaluation. In [26], a method for classification using visual features is presented but its effectiveness is not shown as evaluation is omitted. In contrast, our work considers a large set of images from the popular Web 2.0 site Flickr, allowing to build robust classifiers and ranking models, and combines two different modalities: visual (content-based) and text (meta data) features. In addition, we conduct a large scale evaluation to assess the viability of our approach.

Schmitz *et al.* have formalized folksonomies and discuss the use of association rule mining for analyzing and structuring them in [27]. Work on folksonomy-based web collaboration systems includes [5], [9], and [18] which provide good overviews of social bookmarking tools with special emphasis on folksonomies. A node ranking procedure for folksonomies, the FolkRank algorithm, has been introduced in [11]. FolkRank operates on a tripartite graph of users, resources and items, and generates a ranking of tags for a given user. Another procedure is the Markov Clustering algorithm (MCL) in which a renormalization-like scheme is used in order to detect communities of nodes in weighted networks [29]. A PageRank-like algorithm based on visual links between images is used to improve the ranking function for photo search in [13]. However, none of these these articles are using a combination of community feedback and visual features to classify and rank attractiveness.

There is a plethora of work on classification using a variety of probabilistic and discriminative models [4] and learning regression and ranking functions is well known in the literature [28, 23, 2]. The popular SVM Light software package [15] provides various kinds of parameterizations and variations of SVM training (e.g., binary classification, SVM regression and ranking, transductive SVMs, etc.). In this paper we will apply these techniques, in what is a novel context, to automatic image attractiveness assignment.

To the best of our knowledge, our paper is the first to apply and evaluate automatic classification and ranking methods for photo attractiveness based on visual features and textual meta data. Furthermore, we are the first to propose gathering large training and evaluation sets for photo attractiveness based on community feedback in a Web 2.0 content sharing environment.

3. ATTRACTIVENESS FEATURES FOR IMAGES

Image attractiveness is a very subjective concept influenced by a wide number of factors. In previous studies, it has been shown that high level semantic attributes, such as people expressions or picture composition, are the most relevant when determining the overall appeal of a photo [24]. The current limitations in semantic understanding of images prevent automatic methods from taking advantage of them for the establishment of models. However, there are a number of other attributes which also influence the perception of image attractiveness and that can be measured. This is illustrated in Figure 1, where pairs of semantically affine pictures with varying appeal levels are depicted, showing how semantic properties could just be insufficient for the correct classification of pictures in terms of their attractiveness. In this section we introduce image features available from the content and its associated meta data that we use later for the training of models for image attractiveness classification.

3.1 Visual Features

It is widely accepted that human perception of images is mainly influenced by two factors, namely color distribution and coarseness of the patterns contained [12]. These are complex concepts which convey multiple orthogonal aspects that have to be considered individually. Figure 1 shows several examples. For the same semantic concepts (columns in the figure), very different perceptions of image quality can be perceived. The images in the upper row are generally perceived as more appealing, mainly because of their higher

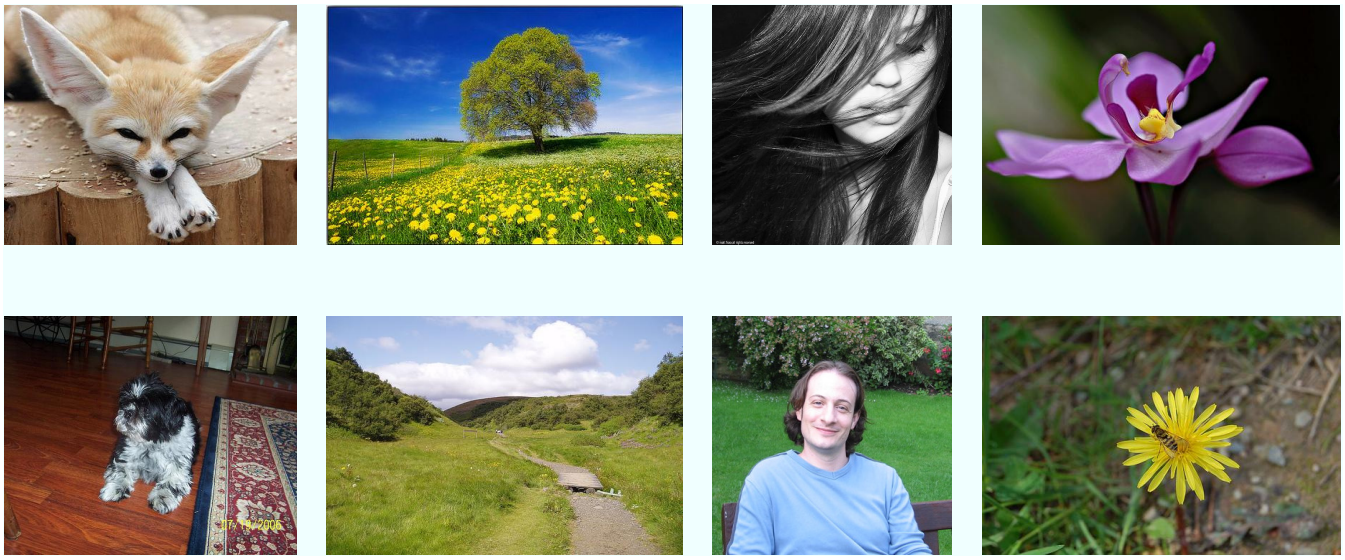

Figure 1: "Attractive" (upper row) vs. "Unattractive" (lower row) images: Each column represents the same semantic concept (animal, landscape, portrait, flower) but differences in appeal-related visual attributes.

artistic value. Even though artistic quality cannot be quantitatively computed, it correlates to certain visual features of images, assigning more optimal values to them. For instance, appealing images tend to have higher colorfulness (column 2), increased contrast (column 3) and sharpness. In this section, we review some of the most relevant visual aspects which we intend to use as image quality indicators.

3.1.1 Color

Color is the pillar of the human vision system. It can be expressed in absolute terms as coordinates in a specific color space. Different color spaces have been defined to suit the requirements of different color-dependent applications. In this section we make use of some of them to establish color attributes of the image. In addition to the well-known sRGB, we also refer to the HSV (Hue-Saturation-Value) and HSL (Hue-Saturation-Lightness) color spaces, which provide a more intuitive representation of colors for humans [1]. The YUV (Luma-Chrominance) color space is also used as it maps luminance intensity (brightness) directly as the Y coordinate. Finally, the CIE$L^*u^*v^*$ color space [1] is the most comprehensive color model, capable of describing the complete visible spectrum. It provides an interesting color decomposition with two chromaticity components, u and v.

The following attributes are commonly used to characterize the color present in images:

Brightness The brightness of a color is a measure of the amplitude of its light wave, or intensity. Even though it is a very simple attribute, it has been effectively used for filtering poorly exposed photos [25]. For images in the YUV color space, it can be straightforwardly determined as the average of the luminance values, Y, of the complete sequence of pixels,

$$\bar{Y} = \frac{1}{N}\sum_{x,y} Y_{xy} \tag{1}$$

where Y_{xy} denotes the luminance value of pixel (x, y) and N denotes the size of the image.

Saturation: The saturation of a color is a measure of its vividness. It is defined as the difference of intensity of the different light wavelengths that compose the color. In the CIE$L^*u^*v^*$ space, saturation is defined by the expression

$$S_{uv} = 13\sqrt{(u' - u'_0)^2 + (v' - v'_0)^2} \tag{2}$$

where u' and v' are the chromaticity coordinates of the considered color, and u'_0 and v'_0 are the corresponding (u', v') coordinates for the white reference color chosen. In other color spaces, including HSV and HSL, various correlates of saturation are directly mapped into their coordinates. According to the definition of HSV, saturation can be established using

$$S = max(R, G, B) - min(R, G, B) \tag{3}$$

where R, G and B are the coordinates of the color the sRGB color space.

Colorfulness: The colorfulness of a color is a measure of its difference against grey. When considering the pixels of an image altogether, the individual distance between pixel colors is also taken into account. Winkler [31] proposes to compute the colorfulness index using the distribution of chroma values. A more efficient method for images coded in the sRGB color space is described by Hasler [10]. The opponent color space is defined as

$$\begin{aligned} rg &= R - G, \\ yb &= \frac{1}{2}(R + G) - B \end{aligned}$$

and colorfulness can be obtained using

$$Cf = \sigma_{rgyb} + 0.3 \cdot \mu_{rgyb}, \tag{4}$$

$$\sigma_{rgyb} = \sqrt{\sigma^2_{rg} + \sigma^2_{yb}}, \tag{5}$$

$$\mu_{rgyb} = \sqrt{\mu^2_{rg} + \mu^2_{yb}} \tag{6}$$

Naturalness: This highly subjective concept aims at providing a measure of the degree of correspondence between

images and human perception of reality. It condenses many aspects of perceived color attributes, such as colorfulness or dynamic range. Huang *et al.* propose a method to obtain a quantitative value [12]. Considering colors are in the HSL color space, they use pixels with $20 \leq L \leq 80$ and $S > 0.1$. These are grouped according to their hue (H coordinate) value in three sets: 'A - Skin', 'B - Grass' and 'C - Sky'. Average saturation values for each group, μ_S, are used to compute local naturalness indexes using the following expressions:

$$N_{\text{Skin}} = e^{-0.5\left(\frac{\mu_S^A - 0.76}{0.52}\right)^2}, \text{if } 25 \leq hue \leq 70$$

$$N_{\text{Grass}} = e^{-0.5\left(\frac{\mu_S^B - 0.81}{0.53}\right)^2}, \text{if } 95 \leq hue \leq 135$$

$$N_{\text{Sky}} = e^{-0.5\left(\frac{\mu_S^C - 0.43}{0.22}\right)^2}, \text{if } 185 \leq hue \leq 260$$

The final naturalness index is given by the expression:

$$N = \sum_i \omega_{\text{i}} N_{\text{i}}, \quad i \in \{\text{'Skin', 'Grass', 'Sky'}\} \tag{7}$$

where ω_i denotes the proportion of pixels of group i in the image.

Contrast: As introduced above, color perception depends heavily on the relation of local luminance variations to the surrounding luminance. Contrast measures this relative variation of luminance. Multiple definitions for computing the contrast index have been proposed. Weber's definition provides a simple way to obtain contrast for simple periodic patterns as:

$$C^W = \frac{\Delta L}{L} \tag{8}$$

The RMS-contrast is commonly used to determine contrast in a way which allows to be compared between independent images:

$$C^{rms} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{9}$$

3.1.2 Coarseness

Coarseness, on the other hand, represents the degree of detail contained in an image. It mainly depends on the quality of the capturing gear and the photographer, and closely relates to the notions of resolution (number of pixels per inch) and acutance (maximum color change ratio per inch). The most commonly used metric to determine the coarseness of images is sharpness. **Sharpness** measures the clarity and level of detail of an image. Its importance in the final appearance of a photo has been repeatedly emphasized by professional photographers and studies on image appeal [24]. Sharpness can be determined as a function of its Laplacian, normalized by the local average luminance in the surroundings of each pixel:

$$Sh = \sum_{x,y} \frac{L(x,y)}{\mu_{xy}}, \text{ with } L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{10}$$

where μ_{xy} denotes the average luminance around pixel (x,y).

3.2 Text Features

In addition to visual features, the textual annotation of images available in Web 2.0 folksonomies such as Flickr can provide additional clues on the attractiveness of photos. This holds partly due to correlations of topics with appealing image content. As an illustrative example we computed a ranked list of tags from a set of 12,000 photos with more than 5 favorite assignments ("attractive") and another set of the same size containing photos without any favorite assignments ("unattractive"). For ranking the tags, we used the Mutual Information (MI) measure [21, 19] from information theory which can be interpreted as a measure of how much the joint distribution of features X_i (terms in our case) deviate from a hypothetical distribution in which features and categories ("attractive" and "unattractive") are independent of each other.

Table 1 shows the top-50 terms extracted for each category. Obviously many of the "attractive" photos contain nature motives (e.g., *sunset*, *flower*, *animals*), have tags relating to photo technology (*canon*, *nikon*, *hdr*), emphasize artistic aspects and colors (*blackandwhite*, *green*, *red*, etc.) and contain positive statements (*supershot*, *colorphotoaward*). "Unattractive" photos, on the other hand, are often about family occasions (e.g., *birthday*, *wedding*, *family*, *dad*) as well as other private events and activities (*graduation*, *party*, *weekend*, *trip*, *camping*) which are of importance for a small circle of friends and family members but less interesting for a larger community of Flickr users; furthermore, the technical quality of some of these photos might be affected by their amateur character.

4. CLASSIFICATION AND REGRESSION MODELS FOR IMAGE ATTRACTIVENESS

In the previous section, we have seen how feature representations of photos can be obtained using analysis of visual content and textual annotations. In this section, we provide a short review of specific classification and regression techniques known from the machine learning literature, and show how these techniques can be applied to our scenario. We use classification models to automatically categorize photos as attractive or unattractive, and regression models to obtain lists of photos ranked by their attractiveness.

4.1 Classifying Attractiveness

In order to classify photos into categories "attractive" or "unattractive" we use a supervised learning paradigm which is based on training items (photos in our case) that need to be provided for each category. Both training and test items, which are later given to the classifier, are represented as multi dimensional feature vectors. These vectors can be constructed using tf or $tf \cdot idf$ weights of tags and the visual features described in Section 3). Photos labeled as "attractive" or "unattractive" are used to train a classification model, using probabilistic (e.g., Naive Bayes) or discriminative models (e.g., SVMs).

How can we obtain sufficiently large training sets of "attractive" or "unattractive" photos? We are aware that the concept of appeal lies in the eye of the beholder, and is highly subjective and problematic. However, the amount of community feedback in Flickr results in large annotated photo sets which hopefully helps to average out noise in various

Table 1: Top-50 terms according to their MI values for attractive vs. unattractive photos in Flickr

Terms for Attractive photos				
abigfave	green	diamondclassphotographer	clouds	colorphotoaward
naturesfinest	impressedbeauty	sunset	blueribbonwinner	film
nature	red	flowers	pink	sea
flower	sky	art	black	hdr
macro	supershot	light	woman	coolest
blue	aplusphoto	explore	soe	street
bw	canon	flickrdiamond	night	beach
anawesomeshot	water	blackandwhite	landscape	animal
nikon	white	color	bravo	sun
portrait	girl	yellow	superbmasterpiece	garden
Terms for Unattractive photos				
2007	camping	madagascar	memorial	race
wedding	festival	memorialday	matt	mt
graduation	may	prague	pics	dad
2006	canyon	china	vietnam	2
party	ubc	cycling	2003	weekend
trip	ubcaagrad07s	cruise	urlaubvacation	kenya
honeymoon	tour	kollegstufenfahrt	kreuzfahrtcruise	part
vacation	family	birthday	commencement	ian
2005	bbq	drinking	mvmarcopolo	regatta
07	softball	vegas	grand	bermuda

forms and, thus, reflects to a certain degree the "democratic" view of a community. To this end we considered distinct thresholds for the minimum number of favorite assignments *NumFav* for photos; in Section 5 we will see that favorites are highly correlated with other kinds of community feedback such as number of comments or views. Formally, we obtain a set $\{(\vec{p_1}, l_1), \ldots (\vec{p_n}, l_n)\}$ of photo vectors $\vec{p_i}$ labeled by l_i with $l_i = 1$ if $NumFav$ lies above a threshold ("positive" examples), $l_i = -1$ otherwise ("negative" examples).

Linear support vector machines (SVMs) construct a hyperplane $\vec{w} \cdot \vec{x} + b = 0$ that separates the set of positive training examples from a set of negative examples with maximum margin. This training requires solving a quadratic optimization problem whose empirical performance is somewhere between linear and quadratic in the number of training items [3]. In real life, the classes in the training data are not always separable. To handle the general case where a single hyperplane may not be able to correctly separate all training points, *slack* variables are introduced in order to relax the constraints of the optimization problem. For a new, previously unseen, photo $\vec{p}$ the SVM merely needs to test whether it lies on the "positive" side or the "negative" side of the separating hyperplane. The decision simply requires computing a scalar product of the vectors $\vec{w}$ and $\vec{p}$. SVMs have been shown to perform very well for various classification tasks (see, e.g., [7, 14]). Other discriminative classifiers (e.g., based on Fisher discriminants) trade off some accuracy for speed [6], but we restrict ourselves to linear SVMs.

4.2 Regression Models for Attractiveness

To learn a regression model we consider training sets $\{(\vec{p_1}, r_1), \ldots, (\vec{p_n}, r_n)\}$ of photo vectors $\vec{p_i}$ along with relevance values $r_i \in \mathbb{R}$ instead of the category labels used for classification. We are considering the number of favorite assignments $NumFav$ for a photo p_i as relevance value, and feature vector representations of photos as described in the previous subsection on classification.

SV-ϵ regression [28] computes a function $f(\vec{x})$ that has a deviation $\leq \epsilon$ from the target relevance values r_i of the training data with a minimum value for ϵ and at the same time is as "flat" as possible. For a family of linear functions $\vec{w} \cdot \vec{x} + b$ "flatness" means that $||\vec{w}||$ is minimized which results in the following optimization problem:

$$\text{minimize } \frac{1}{2}||\vec{w}||^2 \tag{11}$$

$$\text{subject to } \begin{cases} r_i - \vec{w}\vec{p_i} - b \leq \epsilon \\ \vec{w}\vec{p_i} + b - r_i \leq \epsilon \end{cases} \tag{12}$$

Similar to the classification scenario, slack variables can be introduced if the constraints of the optimization problem cannot be met. By means of the learned regression function f, relevance values $f(\vec{p})$ can be assigned to vector representations $\vec{p}$ of new test photos, resulting in a list of photos ranked according to their attractiveness.

5. EXPERIMENTS

In this section, we present the results of our evaluation for automatic detection of photo attractiveness. First, we describe our strategy for gathering a photo collection from Flickr, and elaborate on the characteristics of our data set. Then, we present the outcome of our two-fold evaluation methodology: 1) We examine the influence of the enhanced photo representations on automatic *classification* of photo attractiveness. 2) We apply regression models to obtain *rankings* of photos according to their attractiveness.

5.1 Data

We gathered a sample of photos from Flickr uploaded in the time between June 1 and 7, 2007. We used the Flickr API to query for photos uploaded in 20 minutes time intervals. In this way, we obtained a total of 2.2 M photos in medium size from 185 k users (note that this is just the subset of photos provided by the Flickr API, the actual amount of uploaded photos during that time is larger).

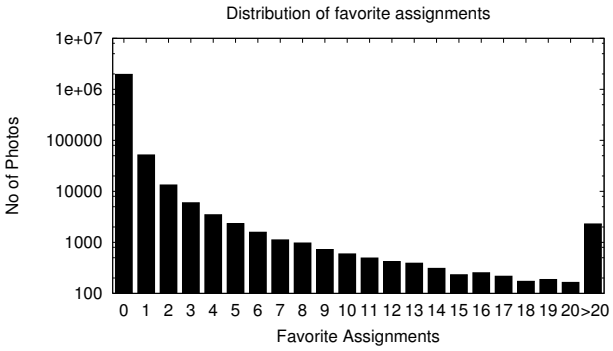

Figure 2: Distribution of favorite assignments

The relatively short time frame of one week (compared to the existence of Flickr) guarantees that for all photos, there was roughly the same chance to obtain community feedback. For each photo, we extracted the number of times the photo was assigned to favorite lists. Figure 2 shows the distribution of the number favorite assignments for these photos. Since adding a photo to one's favorite list is probably the most direct positive assessment, we used the number of favorites as relevance values for building and testing machine learning models. This is also justified by the high correlation of the number of favorite assignments with other important indicators of community interest. We computed the correlation with the number of views/comments and obtained the following values for Kendall's Tau-b: 0.688 for views, 0.767 for comments.

Positive examples were selected using all the photos with at least 2 favorite assignments. We deliberately dismissed photos with just 1 favorite assignment as they do not provide sufficient evidence of social agreement. This resulted in a set of 35,000 photos. In addition we chose a random sample of 40,000 photos without any favorite assignments as the set of negative examples.

5.2 Results

In Section 3, we have presented different methods for extracting visual features and textual features, resulting in enhanced combined feature representations of photos. Machine learning algorithms described in Section 4 make use of this feature information to generate models, and to automatically organize the data. In this section, we show results for classification as well as ranking.

5.2.1 Classification of Attractive Photos

Classifying data into thematic categories usually follows a supervised learning paradigm and is based on training items that need to be provided for each topic. We used the SVM-light [15] implementation of linear support vector machines (SVMs) with standard parameterization in our experiments, as this has been shown to perform well for various classification tasks (see, e.g.,[8, 14]).

We performed different series of binary classification experiments of Flickr photos into the classes "attractive" and "unattractive". We are aware that the concept of appeal is highly subjective and difficult to capture. However, the large amount of community feedback in Flickr allows for a large scale evaluation which hopefully helps to average out noise and reflects, to a certain degree, the view of the community.

For our classification experiments, we considered different levels of restrictiveness for the class "attractive"; to this end we considered distinct thresholds for the minimum number of favorite assignments for a photos ($NumFav \geq 2$, 5, 10 and 20) to be considered as "attractive"; photos without any favorite assignments were considered to belong to the category "unattractive". We considered different amounts of randomly chosen "attractive" training photos (T = 500, 2000, 8000, 20000) as positive examples (where that number of training photos and at least 1000 test photos where available), and the same amount of randomly chosen "unattractive" photos as negative samples. For testing the models based on these training sets we used the disjoint sets of remaining "attractive" photos with same minimum number of assigned favorites and a randomly selected disjoint subset of negative samples of the same size.

We compared the following methods for producing visual features from photos, introduced in Section 3, and build (1-dimensional) feature vectors for classification:

1. **brightness**: computed using equation (1).
2. **contrast**: computed using equation (9)
3. **RGB contrast**: computed using the straightforward extension of equation (9) into the three-dimensional RGB color space.
4. **saturation**: computed as the average of saturation values across the complete sequence of pixels as defined by equation (3)
5. **saturation variation**: computed as the standard deviation of the distribution of values used for saturation.
6. **colorfulness**: computed using equation (4)
7. **sharpness**: computed using equation (10)
8. **sharpness variation**: computed as the standard deviation of the distribution of values used for sharpness.
9. **naturalness**: computed using equation (7)

In addition, we studied the following higher dimensional combined feature vectors:

1. **text**: feature vectors based on the tag representation of the photos using tf weighting
2. **visual**: 9-dimensional feature vectors combining the visual features described above
3. **text+visual**: combination vector obtained from the textual and visual features

Our quality measures are the precision-recall curves as well as the precision-recall break-even points (BEPs) for these curves (i.e. precision/recall at the point where precision equals recall which is also equal to the F1 measure, the harmonic mean of precision and recall in that case). The results for the BEP values are shown in Tables 2 through 5. The detailed precision-recall curves for the example case of T=8,000 training photos and minimum number of favorite assignments $NumFav$=5 are shown in Figure 3. The main observations are:

- The combination vectors obtained from textual and visual features (**text+visual**) provide the best performance. For instance, the configuration with T=8000 positive/negative training photos and minimum NumFav =5, leads to a BEP of 0.8363. Consistently, similar observations can be made for all examined configurations.
- Attractiveness classification based just on textual features (**text**) performs surprisingly well, e.g., BEP =

0.7843 for T=8000 and NumFav $\geq$ 5. This can be explained by a higher interest in certain topics (e.g. woman, car), topic correlation with high technical quality of the pictures (e.g. nature motives, photos annotated by camera-related terms), or, in some cases, quality assignments in tags (e.g. "awesomephoto").

- Although classification using a combination of all visual features (**visual**) is outperformed by classification with textual features (BEP = 0.6664 for T=8000 and NumFav $\geq$ 5) trading recall against precision still leads to applicable results. For instance, we obtain prec=0.7975 for recall=0.3, and prec=0.8472 for recall=0.1; this is useful for finding candidates of attractive photos in large photo sets. Furthermore, classifiers based on visual features have the additional advantage that they can be applied in a more flexible way and in a broader context, e.g., in the absence of textual annotations or in personal photo collections.

We have also studied each of the visual features individually. As expected, each of these features alone proves less powerful than their combination. BEPs are typically around 0.5; however, the precision-recall curves reveal in most cases a clear increase of precision with decreasing recall and thus show that these features are indeed indicators of photo attractiveness. The much higher performance of the combined visual features indicates more complex patterns and relationships between the visual dimensions. Classification results tend to improve, as expected, with increasing number of training photos. Furthermore, the classification performance increases with higher thresholds for the number of favorite assignments for which a photo is considered as "attractive".

5.2.2 Ranking by Attractiveness

Ranking algorithms order a set of objects, Flickr photos in our case, according to their relevance values. For our experiments we chose SVM Regression using the SVMlight [15] implementation with standard parameterization for regression. For training the regression model, we randomly selected 20,000 photos with more than 2 favorites and the same number of photos with 0 favorites. We tested the model on the remaining (disjoint) set of photos with *NumFav*$\geq$2 and on a disjoint set of the same size containing photos with no favorite assignments.

The list of test photos in descending order of their number of favorite assignments was considered as ground truth for our experiments. We compared the order of the automatically generated rankings using Kendall's Tau-b [17]:

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T_1)(P + Q + T_2)}} \qquad (13)$$

where P is the number of concordant pairs, Q is the number of discordant pairs in the lists, T_1 is the number of pairs tied in the first but not in the second list, and T_2 is the number of pairs tied in the second but not in the first list. Values for τ_b can range from -1 to 1. We have chosen the Tau-b version in order to avoid a systematic advantage of our methods due to many ties produced by the high number of photos with same *numFav* value.

We constructed feature vectors based on tags (**text**), single visual features, all visual features (**visual**) and their combination (**text+visual**) in the same way as for the classification experiments described in the previous Section 5.2.1.

Table 6: Ranking using Regression (Kendall's Tau-b): 40000 training photos

Method	Kendall's Tau-b
brightness	0.0006
contrast	-0.0172
RGB contrast	0.0288
saturation	0.1064
saturation variation	0.0472
colorfulness	-0.0497
sharpness	0.0007
sharpness variation	-0.0914
naturalness	0.0143
text	0.3629
visual	0.2523
text+visual	0.4841

The results of the comparison are shown in Table 6. The main observations are very similar to the ones obtained for the classification scenario:

- The combination vectors obtained from textual and visual features (**text+visual**) provide the best ranking performance ($\tau_b = 0.4841$). This value illustrates a remarkable correlation of our model with the actual community feedback, proving the viability of our proposed multi-modal approach.
- Ranking using a combination of all visual features (**visual**) is outperformed by ranking with textual features. However, ranking with only visual features still produces promising results and can be useful for cases and applications where no or insufficient textual photo annotation is available.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown that community feedback in social sharing systems in combination with a multi-modal image representation based on textual annotation and visual features can be used for automatic assignment of photo attractiveness values. More specifically, in what is a novel approach, we have used favorite assignments in the photo sharing environment Flickr to obtain training data and a ground truth for a community-based notion of "attractiveness". We used textual as well as various visual features for constructing vector representation of photos and for building classification and regression models. Our classification and ranking experiments show the best performance for a hybrid combination of tags and visual information. However, the approach of using only visual features shows applicable results as well, and has the advantage of a higher flexibility in the sense that it can be applied in scenarios where no textual meta annotation is available (e.g. personal photo collections or photos without annotations on the web).

We plan to extend and generalize this work to consider various kinds of resources in folksonomies such as still and moving pictures (Youtube) or text (del.icio.us), and use different content and meta data analysis techniques to obtain appropriate object representations. The extension of this work to moving pictures presents several challenges. Perceived quality in this scenario cannot be directly obtained from the analysis of independent frames, as the inherent redundant nature of videos is used by the human brain to

Table 2: Classification Results (BEP): 500 "attractive"/"unattractive" training photos

Method	NumFav ≥ 2	NumFav ≥ 5	NumFav ≥ 10	NumFav ≥ 20
brightness	0.4901	0.5203	0.5227	0.5239
contrast	0.5056	0.4977	0.4985	0.4931
RGB contrast	0.4886	0.4888	0.4888	0.4768
saturation	0.4341	0.413	0.3976	0.3735
saturation variation	0.4684	0.5413	0.4509	0.4325
colorfulness	0.5339	0.4562	0.4503	0.431
sharpness	0.5047	0.5103	0.4844	0.4925
sharpness variation	0.5455	0.5576	0.4438	0.5708
naturalness	0.4841	0.4801	0.4685	0.4547
text	0.6307	0.691	0.7372	0.7798
visual	0.6084	0.631	0.6386	0.6512
text+visual	0.6884	0.7428	0.7693	0.8097

Table 3: Classification Results (BEP): 2000 "attractive"/"unattractive" training photos

Method	NumFav ≥ 2	NumFav ≥ 5	NumFav ≥ 10
brightness	0.5099	0.5199	0.5197
contrast	0.4944	0.4962	0.5042
RGB contrast	0.4888	0.4876	0.4855
saturation	0.4342	0.5862	0.403
saturation variation	0.4686	0.5389	0.4531
colorfulness	0.4664	0.4551	0.5466
sharpness	0.5052	0.5133	0.486
sharpness variation	0.5455	0.4462	0.4406
naturalness	0.5159	0.4791	0.4662
text	0.6758	0.743	0.787
visual	0.6202	0.6441	0.6546
text+visual	0.7373	0.7902	0.8306

Table 4: Classification Results (BEP): 8000 "attractive"/"unattractive" training photos

Method	NumFav ≥ 2	NumFav ≥ 5
brightness	0.4905	0.4798
contrast	0.4939	0.5103
RGB contrast	0.4874	0.4798
saturation	0.5662	0.5828
saturation variation	0.5309	0.4601
colorfulness	0.4668	0.5441
sharpness	0.5052	0.5071
sharpness variation	0.4552	0.4365
naturalness	0.4852	0.5178
text	0.6992	0.7843
visual	0.6384	0.6664
text+visual	0.7798	0.8363

Table 5: Classification Results (BEP): 20000 "attractive"/"unattractive" training photos

Method	NumFav ≥ 2
brightness	0.5085
contrast	0.5084
RGB contrast	0.5156
saturation	0.5675
saturation variation	0.5301
colorfulness	0.4699
sharpness	0.5042
sharpness variation	0.4527
naturalness	0.485
text	0.7193
visual	0.6491
text+visual	0.793

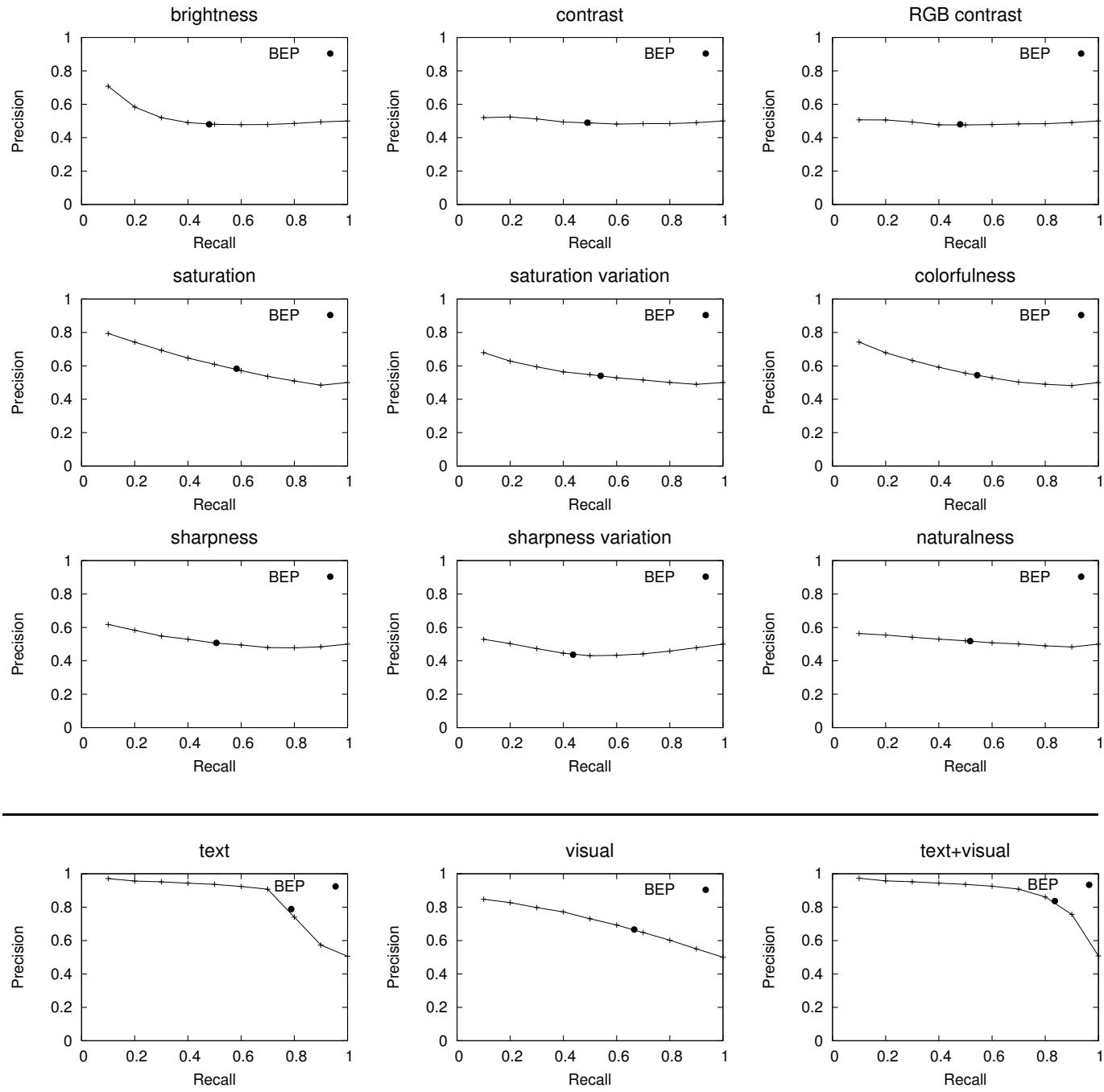

Figure 3: Precision-recall curves for visual and textual dimensions and their combination (8000 training photos per class, numFav$\geq$5)

produce an improved version which is what we ultimately sense. For the case of text documents, besides a simple Bag-of-Words approach which would capture correlations with attractive topics, stylometric features based on sentence structure, vocabulary distributions and linguistic constituents might provide additional clues about the attractiveness of longer texts. Furthermore, we intend to introduce more accurate computations of the described visual attributes, e.g. eigenvalues-based sharpness detection, as well as additional features, such as texture descriptors. In addition, besides an aggregated community-based perception of attractiveness, we would like to study recommender mechanisms taking individual user contexts and preferences into account to provide personalized results.

We think that the proposed techniques have direct applications to search improvement, where automatically computed "attractiveness" can, besides other criteria, be taken into account to improve relevance assignments and rankings.

In this connection, integration and user evaluation within a wider system context and encompassing additional complementary retrieval and mining methods is of high practical importance.

7. ACKNOWLEDGEMENTS

This research was partially funded by the EU Marie Curie ToK Grant *Memoir* (MTKD-CT-2005-030008), and the Large-Scale Integrating EU Project *LivingKnowledge*.

8. REFERENCES

[1] J. Bigun. *Vision with Direction: A Systematic Introduction to Image Processing and Computer Vision*. Springer-Verlag, Secaucus, NJ, USA, 2005.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, New York, NY, USA, 2005. ACM.

[3] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[4] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kauffman, 2002.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proc. 15th Int. WWW Conference*, May 2006.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2000.

[7] S. Dumais and H. Chen. Hierarchical classification of web content. In *SIGIR '00*, pages 256–263, New York, NY, USA, 2000. ACM.

[8] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM'98*, pages 148–155, Maryland, United States, 1998. ACM Press.

[9] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[10] S. Hasler and S. Susstrunk. Measuring colorfulness in real images. volume 5007, pages 87–95, 2003.

[11] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information Retrieval in Folksonomies: Search and Ranking. In *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, 2006. Springer.

[12] 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.

[13] Y. Jing and S. Baluja. Pagerank for product image search. In *WWW*, pages 307–316, New York, NY, USA, 2008. ACM.

[14] T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. *ECML*, 1998.

[15] T. Joachims. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[16] D. Kalenova, P. Toivanen, and V. Bochko. Preferential spectral image quality model. pages 389–398. 2005.

[17] W. H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[18] B. Lund, T. Hammond, M. Flack, and T. Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), 2005.

[19] W. Madison, Y. Yang, and J. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, 1997.

[20] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(8):837–842, 1996.

[21] C. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[22] E. Peli. Contrast in complex images. *Journal of the Optical Society of America*, 7:2032–2040, 1990.

[23] M. Richardson, A. Prakash, and E. Brill. Beyond pagerank: machine learning for static ranking. In *WWW'06*, pages 707–715, NY, USA, 2006. ACM.

[24] A. E. Savakis, S. P. Etz, and A. C. Loui. Evaluation of image appeal in consumer photography. In B. E. Rogowitz and T. N. Pappas, editors, *SPIE Conference Series*, volume 3959, pages 111–120, June 2000.

[25] A. E. Savakis and A. C. Loui. *Method For Automatically Detecting Digital Images that are Undesirable for Placing in Albums*, volume US 6535636. March 2003.

[26] A. E. Savakis and R. Mehrotra. Retrieval and browsing of database images based on image emphasis and appeal. *US 6847733*, 2005.

[27] C. Schmitz, A. Hotho, R. Jaeschke, and G. Stumme. Mining Association Rules in Folksonomies. In *Data Science and Classification*, pages 261–270. Springer Berlin Heidelberg, 2006.

[28] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Kluwer Academic Publishers, 2004.

[29] S. van Dongen. A cluster algorithm for graphs. *National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, Technical Report INS-R0010*, 2000.

[30] C. Y. Wee and R. Paramesran. Measure of image sharpness using eigenvalues. *Inf. Sci.*, 177(12):2533–2552, 2007.

[31] S. Winkler. Visual fidelity and perceived quality: Towards comprehensive metrics. In *in Proc. SPIE*, volume 4299, pages 114–125, 2001.

[32] S. Winkler and C. Faller. Perceived audiovisual quality of low-bitrate multimedia content. *Multimedia, IEEE Transactions on*, 8(5):973–980, 2006.

[33] S. Yao, W. Lin, S. Rahardja, X. Lin, E. P. Ong, Z. K. Lu, and X. K. Yang. Perceived visual quality metric based on error spread and contrast. In *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pages 3793–3796 Vol. 4, 2005.