An Image is Worth More than a Thousand Favorites: Surfacing the Hidden Beauty of Flickr Pictures

Rossano Schifanella
University of Turin
Turin, IT
schifane@di.unito.it

Miriam Redi
Yahoo Labs
Barcelona, SP
redi@yahoo-inc.com

Luca Maria Aiello
Yahoo Labs
Barcelona, SP
alucca@yahoo-inc.com

Abstract

The dynamics of attention in social media tend to obey power laws. Attention concentrates on a relatively small number of popular items and neglecting the vast majority of content produced by the crowd. Although popularity can be an indication of the perceived value of an item within its community, previous research has hinted to the fact that popularity is distinct from intrinsic quality. As a result, content with low visibility but high quality lurks in the tail of the popularity distribution. This phenomenon can be particularly evident in the case of photo-sharing communities, where valuable photographers who are not highly engaged in online social interactions contribute with high-quality pictures that remain unseen. We propose to use a computer vision method to surface beautiful pictures from the immense pool of near-zero-popularity items, and we test it on a large dataset of creative-commons photos on Flickr. By gathering a large crowdsourced ground truth of aesthetics scores for Flickr images, we show that our method retrieves photos whose median perceived beauty score is equal to the most popular ones, and whose average is lower by only 1.5%.

1 Introduction

One of the common uses of online social media surely is to accrue social capital by winning other people's attention (Steinfield, Ellison, and Lampe 2008; Smith and Giraud-Carrier 2010; Burke, Kraut, and Marlow 2011; Bohn et al. 2014). The ever-increasing amount of content produced by the crowd triggers emergent complex dynamics in which different pieces of information have to compete for the limited attention of the audience (Romero et al. 2011). In this process, only few individuals and the content they produce emerge and become popular, while the vast majority of people are bound to a very limited visibility, their contributions being rapidly forgotten (Cha et al. 2007; Sastry 2012). Such dynamics do not necessarily promote high-quality content (Weng et al. 2012), possibly confining some valuable information and expert users in the very tail of the popularity distribution (Goel et al. 2010). This might cause a loss to the community, first because tail contributors are likely to lose engagement and churn out (Karnstedt et al. 2011), but also because tail content is often less curated and difficult to find through search (Baeza-Yates and Sáez-Trumper 2013).

Previous work has focused extensively on studying the patterns of popularity of social media users and of all sorts of online content, trying to isolate the predictive factors of success (Suh et al. 2010; Hong, Dan, and Davison 2011; Brodersen, Scellato, and Wattenhofer 2012; Khosla, Das Sarma, and Hamid 2014). However, considerably less effort has been spent in finding effective ways to surface high-quality content from the sea of forgetfulness of the popularity tail. Finding valuable content in the pool of unpopular items is an intrinsically difficult task because tail items *i)* are many, outnumbering by orders of magnitude those with medium or high popularity, *ii)* have most often low quality, making random sampling strategies substantially ineffective, and *iii)* tend to be less annotated and therefore more difficult to index.

We contribute to tackle these problems in the context of photo sharing services. We use a computer vision method to surface beautiful pictures among those with near-zero-popularity, with no need of additional metadata. Our approach is supervised and relies on features developed in the field of computational aesthetics (Datta et al. 2006). To train our framework, we collect for the first time a large ground truth of aesthetic scores assigned to Flickr images by non-expert subjects via crowdsourcing. Differently from conventional aesthetics datasets (Datta et al. 2006; Murray, Marchesotti, and Perronnin 2012), our ground truth includes images with a wide spectrum of quality levels and better reflects the taste of a non-professional public, making it the ideal training set to classify web images.

When tested on nearly 9M creative-commons Flickr pictures, our method is able to surface from the set of photos that received very low attention ($\leq$5 favorites) a selection of images whose perceived beauty is close to that of the most favorited ones, with the same median value and an average value that is just 1.5% lower. Results are consistent for images in four different topical categories and largely outperform a random baseline, computer vision methods trained on traditional aesthetics databases, and a state-of-the-art computer vision methods targeted to the prediction of image popularity (Khosla, Das Sarma, and Hamid 2014).

We summarize our main contributions as follows:

- We build and make publicly available[1] the largest ground truth of aesthetic scores for Flickr photos constructed so far, including 10.800 pictures of 4 different topical categories and 60K judgments. We carefully designed the crowdsourcing experiment to account for the biases that can incur in a task that is characterized by a strong subjective component.
- We provide an analysis of ordinary people's aesthetics perception of web images. We find that perceived beauty and popularity are correlated ($\rho = 0.43$) but the beauty scores of very popular items have higher variance than unpopular ones. We find that a non-negligible number of unpopular items are extraordinarily appealing.
- We propose a method to retrieve beautiful yet unpopular images from very large photo collections. Our approach works in a cold start scenario as it needs in input only the visual information of the picture. Also, it overcomes the issue of sparsity (i.e., few beautiful pictures hidden among very large amounts of mediocre images) with surprisingly high precision, being able to retrieve images whose perceived beauty is comparable to the top-rated photos.

After a review of the related work (§2), we touch upon the popularity skew in Flickr (§3). We then describe the process of collection of the aesthetics scores through crowdsourcing (§4). Next, we describe the computer vision method we use to identify beautiful pictures (§5) and we report the aesthetic prediction results in comparison with other baselines (§6). Last, we show that our method can surface beautiful photos from a large pool of non-popular ones (§7).

2 Related work

Popularity Prediction. Being able to characterize and predict item popularity in social media is an important, yet not fully solved task (Hong, Dan, and Davison 2011). The possibility of predicting the popularity of videos and pictures in social platforms like YouTube, Vimeo, and Flickr has been explored extensively (Cha, Mislove, and Gummadi 2009; Figueiredo, Benevenuto, and Almeida 2011; Brodersen, Scellato, and Wattenhofer 2012; Ahmed et al. 2013). Multimodal supervised approaches that combine metadata and computer vision features have been used to predict photo popularity. Visual features like coarseness and colorfulness, well predict the number of favorites in Flickr (San Pedro and Siersdorfer 2009) and the number of reshares in Pinterest to some extent (Totti et al. 2014). The presence of specific visual concepts in the image, such as human faces (Bakhshi, Shamma, and Gilbert 2014), are good predictors too. Recently, Khosla et al. (Khosla, Das Sarma, and Hamid 2014) have made one of the most mature contributions in this area, training a SVR model on both visual content and social cues to predict the normalized view count on a large corpus of Flickr images. While previous work tries to understand why popular images are successful, we flip the perspective to see if high-quality pictures hide in the long tail and to what extent we are able to automatically surface them. This necessity is also supported by the weak correlation between received attention and perceived quality found in small image datasets (Hsieh, Hsu, and Wang 2014).

Popularity vs. Quality. Both social and computer scientist have investigated the relation between popularity and intrinsic quality of content. Items' popularity is only partly determined by their quality and it is largely steered by the early popularity distribution, often with unpredictable patterns (Salganik, Dodds, and Watts 2006). User's limited attention drives the popularity persistence and virality of an item more than its intrinsic appeal (Weng et al. 2012; Hodas and Lerman 2012). A piece of content can attract attention because of many factors including the favorable structural position of its creator in a social network (Hong, Dan, and Davison 2011), the sentiment conveyed by the message (Quercia et al. 2011), or the demographic (Suh et al. 2010) and geographic (Brodersen, Scellato, and Wattenhofer 2012) composition of the audience. On video (Sastry 2012) or image (Zhong et al. 2013) sharing platforms, the content that receives larger shares of attention is often of niche topical interest. Adopting community-specific behavioural norms can also increase popularity returns. On Twitter, users who generate viral posts are those who limit their tweets to a single topic (Cha et al. 2010). On Facebook, communicating along weak ties is the key to spread content (Bakshy et al. 2012). More in general, social activity, even in its most superficial meaning (e.g., "poking") can be a powerful attractor of popularity (Vaca Ruiz, Aiello, and Jaimes 2014; Aiello et al. 2012).

Computational Aesthetics. Computational aesthetics is the branch of computer vision that studies how to automatically score images in terms of their photographic beauty. Datta et al. (2006) and Ke et al. (2006) designed the first compositional features to distinguish amateur from professional photos. Computational aesthetics researchers have been developing dedicated discriminative visual features and attributes (Nishiyama et al. 2011; Dhar, Ordonez, and Berg 2011), generic semantic features (Marchesotti et al. 2011; Murray, Marchesotti, and Perronnin 2012), topic-specific models (Luo and Tang 2008; Obrador et al. 2009) and effective learning frameworks (Wu, Hu, and Gao 2011) to improve the quality of the aesthetics predictors. Aesthetic features have been also used to infer higher-level properties of images and videos, such as image affective value (Machajdik and Hanbury 2010), image memorability (Isola et al. 2011), video creativity (Redi et al. 2014b), and video interestingness (Redi and Merialdo 2012; Jiang et al. 2013). To our knowledge, this is the first time that image aesthetic predictors are used to expose high quality content from low-popular images in the context of social media.

Ground Truth for Image Aesthetics. Existing aesthetic ground truths are often derived from photo contest websites, such as DPChallenge.com (Ke, Tang, and Jing 2006) or Photo.net (Datta et al. 2006), where (semi) professional photographers can rate the quality of their peers' images. The average quality and style of the images in such datasets is way higher than the typical picture quality in photo shar-

[1] http://di.unito.it/beautyicwsm15

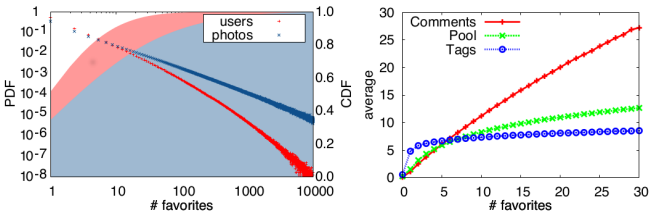

Figure 1: (Left) Distribution of the number of favorites for Flickr photos and users. (Right) Average number of comments, tags, and uploads to group photo pools for photos with a fixed number of favorites.

ing sites, making them not suitable to train *general* aesthetic models. Hybrid datasets (Luo, Wang, and Tang 2011) that add lower-quality images to overcome this issue are also not good for training (Murray, Marchesotti, and Perronnin 2012). In addition, social signals such as Flickr interestingness[2] (Jiang et al. 2013) are often used as a proxy for aesthetics in that type of datasets. However, no quantitative evidence is given that neither the Flickr interestingness nor the popularity of the photographers are good proxies for image quality, which is exactly the research question we address. Crowdsourcing constitutes a reliable way to collect ground truths on image features (Redi and Povoa 2014), the only attempt to do it in the context of aesthetics has been limited in scope (faces) and very small-scale (Li et al. 2010).

3 Popularity in Flickr

Flickr is a popular social platform for image sharing. Users can establish directed *social links* by "following" other users to get updates on their activity. Users can label their own photos with free-text *tags* and publish them in the photo pools of *groups*. Every public photo can be marked as *favorite* or annotated with a textual *comment* by any user in the platform. Flickr also maintains and updates periodically the *Explore* page[3], a showcase of interesting photos.

The complex dynamics that attract attention towards Flickr images revolve around all the above mentioned mechanisms of social feedback that, as in any other social network, tend to promote some items more than others. As a result, the distribution of picture popularity —usually measured by the number of favorites (Cha, Mislove, and Gummadi 2009)— is very broad. Figure 1 (left) shows statistics on user and image popularity computed over a random sample of 200M public Flickr photos that have been favorited at least once. The distribution of the mass of favorites over the photos is highly unequal (Gini coefficient 0.68): the number of favorites of the pictures in this sample spans four orders of magnitude, with the majority of them having only one favorite (52%). The same figure holds when aggregating the popularity by users: some accumulate thousands favorites while the vast majority (∼70%) rustles up less than ten.

As for the intuition given by the *Infinite Monkey Theorem*, the unpopular users must be able to collectively produce a certain amount of exceptionally valuable content just because of their substantial number. More concretely, it is hard to believe that there is no high-quality photo among 166M pictures with five favorites or less. Estimating how many beautiful pictures lie in the popularity tail and understanding how we can draw those out of the mass of user-generated content are the main goals of this contribution.

One may think that one possibility to achieve the goal would be to leverage different types of social feedback (e.g., comment). However, unpopular items rarely receive social feedback. As displayed in Figure 1 (right), the number of comments, tags, and uploads in groups is positively correlated with the number of favorites, with near-zero favorite pictures having a near-zero amount of all the other metrics, on average. Providing a method that does not rely on any type of explicit feedback has therefore the advantage of being more general and suitable for a cold-start scenario. For this reason, we rely on a supervised computer vision method that we describe in §5 and whose training set is collected as described in the next section.

Category	Tags
people	people, face, portrait, groupshot
nature	flower, plant, tree, grass, meadow, mountain
animals	animal, insect, pet, canine, carnivore, butterfly, feline, bird, dog, peacock, bee, lion, cat
urban	building, architecture, street, house, city, church, ceiling, cityscape, brick, tower, window, highway, bridge

Table 1: Set of machine tags included in each image category

4 Ground truth for image aesthetics

We build a ground truth for aesthetics from a 9M random sample of the Creative Commons Flickr Images dataset[4]. We collect the annotations using CrowdFlower[5], a large crowdsourcing platform that distributes small, discrete *tasks* to online *contributors*. Next we describe how we selected the images for our corpus (§4.1), how we run the crowdsource experiment (§4.2), and the results on the beauty judgments we got from it (§4.3).

4.1 Definition of the image corpus

To help the contributor in the assessment of the image beauty, we build a photo collection that *i)* presents topically coherent images and *ii)* represents the full popularity spectrum, thus ensuring a diverse range of aesthetic values.

Topical Coherence. Different picture categories can achieve the same aesthetic quality driven by different criteria (Luo, Wang, and Tang 2011). To make sure that contributors use the same evaluation standard, we group the images in classes of coherent subject *categories*. To do that, we use Flickr *machine tags*[6], namely tags assigned by a computer vision classifier trained to recognize the type of subject depicted in a photo (e.g., a bird or a tree) with a certain confidence level. We manually group the most frequent machine

[2] Flickr interestingness algorithm is secret, but it considers some metrics of social feedback. For more details refer to https://www.flickr.com/explore/interesting

[3] https://www.flickr.com/explore

[4] http://bit.ly/yfcc100m

[5] http://www.crowdflower.com

[6] http://bit.ly/1umsOnL

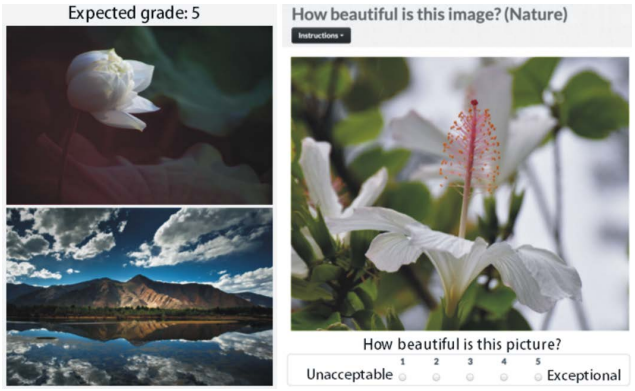

Figure 2: Screenshot of the crowdflower job: instruction examples (left) and voting task (right).

tags in topically-coherent macro-groups, coming up with 4 final categories: *people*, *nature*, *animals*, and *urban*. We only consider the pictures associated with high-confidence machine tags ($\geq$0.9). Moreover, we manually clean the final photo selection by replacing few instances that suffered from machine tag misclassification. The full list of machine tags per category is reported in Table 1.

Full Popularity Range. Within each category, we are interested in assessing the perceived beauty of photos with different popularity levels. To do so, we identify three popularity buckets obtained by logarithmic binning over the range of number of favorites f. We refer to them as *tail* ($f \leq 5$), *torso* ($5 < f \leq 45$), and *head* ($f > 45$). The tail of the distribution contains 98% of the photos, whereas the torso and head contain 1.6% and 0.4% respectively. We randomly sample, within each category, 1000 images from the *tail* and 1000 from the *torso*. Because of the reduced number of most popular pictures we do not sample randomly the *head* bucket but we consider the top 500 instead. Images from such diverse popularity levels are likely to take a wide range of aesthetic values, thus ensuring diversity in our corpus, very important to get reliable beauty judgements (Redi et al. 2014a).

4.2 CrowdFlower experiment

Crowdsourcing tasks are influenced by a variety of human factors that are not always easy to control (Mason and Suri 2012). However, platforms like CrowdFlower offer advanced mechanisms to tune the annotation process and enable the best conditions to get high-quality judgments. To facilitate the reproducibility of our experiment, next we report the main setup parameters.

Task interface and setup. The task consists in looking at a number of images and evaluating their aesthetic quality. At the top of the page we report a short description of the task and we ask *"How beautiful is this picture?"*. The contributor is invited to judge the intrinsic beauty of an image and not the appeal of its subject; high quality, artistic pictures that depict a non-conventionally beautiful subject (e.g., a spider), should be marked as beautiful and viceversa. Screenshots of the Crowdflower job interface are shown in Figure 2.

Although several approaches and rating scales can be used to get quality feedback (Fu et al. 2014), we use the 5-point *Absolute Category Rating* (ACR) scale, ranked from *"Unacceptable"* to *"Exceptional"*, as it is a good way to collect aesthetic preferences (Siahaan, Redi, and Hanjalic 2013). To help the annotators in their assessment, two example images and a textual description of each grade are provided (see Figure 2 and Table 2). The examples are Flickr images that have been unanimously judged by three independent annotators to be clear representatives of that beauty grade. Below the examples, each page contains 5 randomly selected images (*units* of work in CrowdFlower jargon), each followed by the radio buttons to cast the vote. The random selection of images allows us to mix pictures from different popularity ranges in the same page, thus offering to the users an easier context for comparison (Fu et al. 2014). We show all the images with approximately the same (large) size because image size can skew the perception of image quality (Chu, Chen, and Chen 2013).

1	**Unacceptable**	Extremely low quality, out of focus, underexposed, badly framed images
2	**Flawed**	Low quality images with some technical flaws (slightly blurred, slightly over/underexposed, incorrectly framed) and without any artistic value
3	**Ordinary**	Standard quality images without technical flaws (subject well framed, in focus, and easily recognizable) and without any artistic value
4	**Professional**	Professional-quality images (flawless framing, focus, and lightning) or with some artistic value
5	**Exceptional**	Very appealing images, showing both outstanding professional quality (photographic and/or editing & techniques) and high artistic value

Table 2: Description of the five-level aesthetic judgment scale

Each photo receives at least 5 judgments, each one by an independent contributor. Each contributor can submit a maximum of 500 judgments, to prevent a predominance of a small group of workers. Contributors are geographically limited to a set of specific countries[7], to ensure higher cultural homogeneity in the assessment of image aesthetics (Hagen and Jones 1978). Only contributors with an excellent track record on the platform (responsible for the 7% of monthly CrowdFlower judgments overall) have been allowed. We also banned workers that come from external crowdsourcing channels that have a ratio of trusted/untrusted users lower than 0.9.

Quality control. *Test Questions* (also called *Gold Standard*) are used to test and track the contributor's performance and filter out bots or unreliable contributors. To access the task, workers are first asked to annotate correctly 6 out of 8 *Test Questions* in an initial *Quiz Mode* screen and their performance is tracked throughout the task with *Test Questions* randomly inserted in every task, disguised as normal units. To support the learning process of a contributor, we tag each

[7] Australia, Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Poland, Spain, Sweden, United Kingdom, United States

	Units	Judgments	Workers	Countries	Trust
people	2500	12725	141	13	0.843
nature	2500	15054	178	14	0.841
animals	2500	13269	117	13	0.80
urban	2500	13213	111	13	0.839

Table 3: General statistics on the crowdsourcing experiment

Test Question with an explanation that pops up in case of misjudgment (e.g., "excellent combination of framing, lightning, and colors resulting in an artistic image, visually very appealing" is one of the description for an high rated item).

To build the set of *Test Questions*, we first collected about 200 candidate images from different online sources including Flickr, web repositories, aesthetics corpora (Murray, Marchesotti, and Perronnin 2012), and relevant photos retrieved by the main image search engines. Three independent editors manually annotated the candidate sets with a beauty score. For each category, we run a small-scale pilot CrowdFlower experiment to consolidate the editors' assessment taking into account the micro-workers feedback. This process led us to mark some of the *Test Question* with two contiguous scores. After this validation step, we identified for each grade the set of images with the highest inter-rater agreement for a total of 100 images.

4.3 Results

We run a separate job for each topical category. Table 3 summarizes the number of units annotated, judgments submitted, distinct participants, and the average accuracy (trust) on *Test Questions* of the contributors. Each unit can receive more than 5 independent judgments; in the case of *nature* we collected 20% more judgments than for the other categories. On average, more than 140 contributors geographically distributed in 13 countries and characterized by a high level of trustworthiness participated to each experiment.

Inter-rater agreement. To assess the quality of the collected data, we measure the level of agreement between annotators. Table 4 shows a set of standard measures to evaluate the inter-rater consistency. *Matching%* is the percentage of matching judgments per item. Across categories the agreement is solid, with an average of 70%. However, the ratio of matching grades does not capture entirely the extent to which agreement emerges. In fact, the task is inherently subjective and in some cases the quality of an image naturally converges to an intermediate level. We therefore compute the Fleiss' K, a statistical measure for assessing the reliability of the agreement between a fixed number of raters. Since Fleiss' K is used to evaluate agreements on categorical ratings, it is not directly applicable to our task. We therefore binarize the task, and assign to each judgment either a $Beautiful$ or $NotBeautiful$ label, according to the score being respectively greater or lower than the median. Consistently, the Fleiss' K shows a fair level of agreement. To further evaluate inter-participant consistency we computed the Cronbach's α that has been extensively adopted in the context of assessing inter-rater agreement on aesthetics tasks (Siahaan, Redi, and Hanjalic 2013). For all categories, the Cronbach's coefficient lies in the interval $0.7 \leq \alpha < 0.9$ that is commonly defined as a $Good$ level of consistency.

	Matching%	Fleiss' K	Cronbach's α
people	68.82	0.38	0.74
nature	72.65	0.27	0.71
animals	69.37	0.35	0.8
urban	73.13	0.38	0.8

Table 4: Measures of judgment agreement

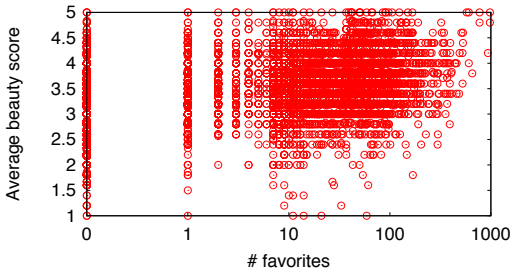

Figure 3: Relation between popularity (number of favorites) and crowdsourced beauty scores for 10,800 Flickr pictures.

Beauty judgements. The Spearman correlation ρ between the number of favorites and the average beauty score is 0.43. Although the correlation is substantial, the variability of perceived beauty for each popularity value is very high. In Figure 3 we plot the beauty score against the number of favorites, for each photo. Zero-popularity images span the whole aesthetics judgment scale, from 1 to 5, and most popularity levels have photos within the $[2.5, 5]$ beauty range. Very low scores (1,2) are rare. This picture confirms our initial motivation as it shows instances of unpopular yet beautiful photos, as well as a good portion of very popular photos with average or low quality.

Results on the distribution of judgments across categories and popularity buckets are summarized in Figure 4. As expected, the $high$ bucket shows the highest average score followed by the $medium$ and the low. With the exception of the $people$ category, the standard deviation follows the same trend: higher popularity corresponds to higher disagreement. This might be due to the fact that viewers are likely to largely agree on objective elements that make an image non-appealing, such as technical flaws (e.g., bad focus) but on the other hand they might not agree on what makes an image exceptionally beautiful, which can be a more subjective characteristic. Given that the more a photo is popular the more it tends to be appealing, this phenomenon can partly explain the inconsistent agreement level among popularity buckets. Across categories we observe that $animals$ images have the highest average quality perception (3.49 ± 0.75) while the remaining categories show a mean around 3.31.

5 Image Aesthetics

Having collected a ground truth of crowdsourced beauty judgements, we now design a computational aesthetic framework to surface beautiful, unpopular pictures. Our method is based on regressed *compositional features*, namely visual features that are specifically designed to describe how much an image fulfills standard photographic rules. We design our framework as follows:

Visual Features. We design a set of visual features to ex-

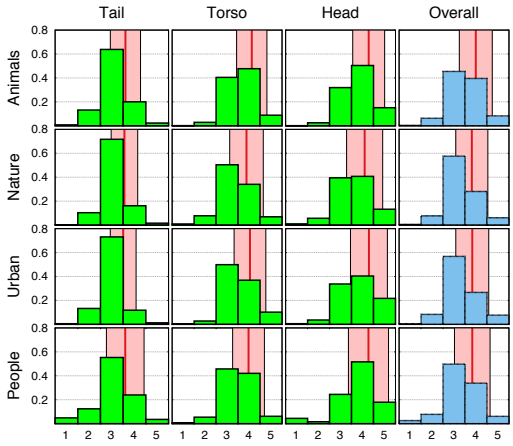

Figure 4: Distribution of ratings across categories and popularity buckets. The red lines and their surrounding areas represent the average and standard deviation.

pose image photographic properties. More specifically, we compose a 47-dimensional feature vector with the following descriptors:

- *Color Features.* Color patterns are important cues to understand the aesthetic and affective value of a picture. First, we compute a *Contrast* metric, that provides information about the distinguishability of colors based on the magnitude of the average luminance:

$$Contrast = \frac{Y_{max} - Y_{min}}{\bar{Y}} \quad (1)$$

where $Y_{max}, Y_{min}, \bar{Y}$ correspond respectively to maximum, minimum, and average of the luminance channel.

We then extract the average of the *Hue, Saturation, Brightness (H,S,V)* channels, computed both on the whole image and on the inner quadrant resulting after a 3x3 division of the image, similar to previous approaches (Datta et al. 2006). By combining average Saturation ($\bar{S}$) and Brightness ($\bar{V}$) values, we also extract three indicators of emotional dimensions, *Pleasure, Arousal and Dominance*, as suggested by previous work on affective image analysis (Machajdik and Hanbury 2010):

$$\begin{aligned} Pleasure &= 0.69\bar{V} + 0.22\bar{S} \\ Arousal &= -0.31\bar{V} + 0.60\bar{S} \\ Dominance &= 0.76\bar{V} + 0.32\bar{S} \end{aligned} \quad (2)$$

After quantizing the HSV values, we also collect the occurrences of 12 Hue bins, 5 Saturation bins, and 3 Brightness bins in the HSV *Itten Color Histograms*. Finally, we compute *Itten Color Contrasts*, i.e. the standard deviation of H, S and V Itten Color Histograms (Machajdik and Hanbury 2010).

- *Spatial Arrangement Features.* Spatial arrangement of objects, shapes and people plays a key role in the shooting of good photographs (Freeman 2007). To analyze the spatial layout in the scene, first, we resize the image to a squared matrix $\mathbf{I_{ij}}$, and we compute a *Symmetry* descriptor based on the difference of the Histograms of Oriented Gradients (HOG) (Dalal and Triggs 2005) between the left half of the image and its flipped right half:

$$Symmetry = ||\Phi(\mathbf{I^l}) - \Phi((\mathbf{I} \cdot \mathbf{J})^\mathbf{r})||_2, \quad (3)$$

where Φ is the HOG operation, $\mathbf{I^l}$ is the left half of the image, and $(\mathbf{I} \cdot \mathbf{J})^\mathbf{r}$ is the flipped right half of the image, being $\mathbf{J}$ the anti diagonal identity matrix that imposes the left-right flipping of the columns in $\mathbf{I_{ij}}$. We also consider the *Rule of Thirds*, a photographic guideline stating that the important compositional elements of a picture should lie on four ideal lines (two horizontal, two vertical) that divide it into nine equal parts (the thirds). To model it, from the resized image $\mathbf{I_{ij}}$, we compute the a saliency matrix (Hou and Zhang 2007), exposing the image regions that are more likely to grasp the attention of the human eye. We then analyze the distribution of the salient zones across the image thirds by retaining the average saliency value for each third subregion.

- *Texture Features.* We describe the overall complexity and homogeneity of an image by computing the Haralick's features (Haralick 1979), namely the *Entropy, Energy, Homogeneity, Contrast* of the Gray-Level Co-occurrence Matrices.

Groundtruth. We use our crowdsourced groundtruth as the main source of knowledge for our supervised framework. Since topic-specific aesthetic models have been shown to perform better than general frameworks (Luo, Wang, and Tang 2011), we keep the division of the ground truth into semantic categories (*people*, *urban*, *animals*, *nature*), and learn a separate, topic-specific aesthetic model for each of them.

Learning Framework. We train category-specific models using *Partial Least Squares Regression (PLSR)*, a very effecive prediction framework for visual pattern analysis (McIntosh et al. 1996). For each semantic category, PLSR learns a set of *regression coefficients*, one per dimension of the visual feature vector, by combining principles of least-squares regression and principal component analysis. Each category-specific group of regression coefficients constitutes a separate aesthetic model.

Prediction and Surfacing. We apply the models to automatically assess the aesthetic value of new, unseen images (i.e., images that do not belong to the training set). To do so, we use the regression coefficients in a linear combination with the features of each image, thus obtaining the predicted aesthetic score for that image.

We use our aesthetic models for two types of experiments. First, to study the performance of our framework against similar approaches, we run a small-scale experiment where the task is to *predict* the aesthetic scores of the crowdsourced groundtruth. We then apply the aesthetic models to *rank* a very large set of images in terms of beauty, with the aim of surfacing the most appealing non-popular pictures.

6 Beauty Prediction from and for the Crowd

To test the power of our aesthetics predictor, we run a small-scale experiment on the crowd-sourced dataset. We look at

	CrowdBeauty	MIT popularity	TraditionalBeauty	Random
animals	0.54	0.37	0.251	0.001
urban	0.46	0.27	0.12	0.003
nature	0.34	0.29	0.11	-0.003
people	0.42	0.31	0.27	-0.008

Table 5: Spearman correlation between the crowdsourced beauty judgments and the scores given by different methods on the images of the test set.

how much the aesthetic scores assigned by our framework correlate with the actual beauty scores assigned by the workers, and evaluate the performance of our algorithm against other ranking strategies.

Baselines. We compare our method with two baselines:

Popularity Predictor: What if a popularity predictor was enough to assess image beauty? To check that, we compare our algorithm with an established content-based image popularity predictor. For each picture in our ground truth, we query the MIT popularity API[8], a recently proposed framework that automatically predicts image popularity scores (in terms of normalized view count) score given visual cues, such as colors and deep learning features (Khosla, Das Sarma, and Hamid 2014).

Traditional Aesthetic Predictor: What if existing aesthetic frameworks were general enough to assess crowdsourced beauty? As mentioned in §5, our models are specifically trained on the crowdsourced dataset, i.e., a groundtruth of images generated and voted by average users. On the other hand, existing aesthetic predictors are generally trained on semi-professional images evaluated by professional photographers. To justify our dataset collection effort, we show how a classifier trained on traditional aesthetic datasets performs in comparison with our method. We design this baseline with the same structure and features as our proposed method, but, instead of using our crowdsourced ground truth, we train on the AVA dataset (Murray, Marchesotti, and Perronnin 2012). Similar to our method, we build one category-specific model for each semantic category. This is achieved by training each category-specific model with the subset of AVA pictures in the corresponding category. We infer the category according to tags attached to each image, as proposed for many topic-specific aesthetic models (Luo and Tang 2008; Obrador et al. 2009).

Experimental Setup. To evaluate our framework, for each semantic category we retain 800 images for test and the rest for training. For training, we use images from all the 3 popularity ranges (tail, torso, head). For test, we consider non-popular images only, as our main purpose is to detect "hidden" beautiful pictures with low number of favorites. For both training and test, we use the total of 47 visual features, that are reduced to 15 components by the PLSR algorithm.

We then score the images in the test set using the output of our framework, the MIT popularity scores, the output of the traditional aesthetic classifier, and a random baseline. Next, we evaluate the performance of the three algorithms in terms of Spearman Correlation Coefficient between the scores predicted on the test set by each model, and the actual votes from the crowd. This metric gauges the ability of each model to replicate the human aesthetic preferences on non-popular Flickr images.

[8] http://popularity.csail.mit.edu

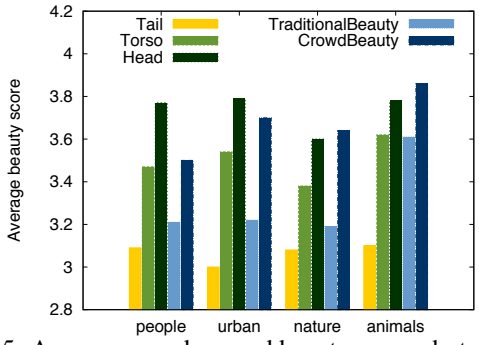

Figure 5: Average crowdsourced beauty score photos in different popularity buckets and for photos surfaced by the aesthetics predictors.

Experimental Results. The correlation between the beauty scores assigned by the micro-workers on the test set and our proposed algorithm (*CrowdBeauty* in the following) is substantially high for all categories, as shown in Table 5. In particular, the most predictable class is the *animals* category, followed by *urban*. The higher performance in these two cases might be due to the smaller range of poses and compositional layouts available to the photographer when shooting pictures of subjects belonging to these particular categories. As expected, the results of the random approach are completely uncorrelated from the beauty scores. For all semantic categories, we see that our method outperforms both the popularity predictor (*MIT Popularity*) and the traditional aesthetic classifier (*TraditionalBeauty*), showing the usefulness of building a dedicated ground truth and aesthetic classifier to score non-popular web images.

7 Surfacing Beautiful Hidden Photos

Having provided some evidence about the effectiveness of our approach, we apply it in a more realistic scenario where the goal is to surface beautiful images from a large number of non-popular Flickr pictures.

To do so, we compute the features described in §5 on all the 9M images of the large-scale categorized dataset of creative commons Flickr images in our dataset. We apply the category-specific model on the pictures in each topical category separately and rank the pictures by their predicted aesthetics scores. For the sake of comparison, we repeat the same procedure with the traditional aesthetic models (*TraditionalBeauty*) used as baseline in §6, and rank them in terms of the predicted beauty scores. We do not consider here the MIT Popularity baseline as its scores can only be retrieved via API with a certain request delay, which it is not practical for a very large set of images.

To quantify how appealing the images surfaced with our approach are, we implemented an additional crowdsourcing experiment in which images with different popularity levels are evaluated against the top-ranked images according to our models and the traditional aesthetic model. We replicated the same experimental settings described in Section 4 and

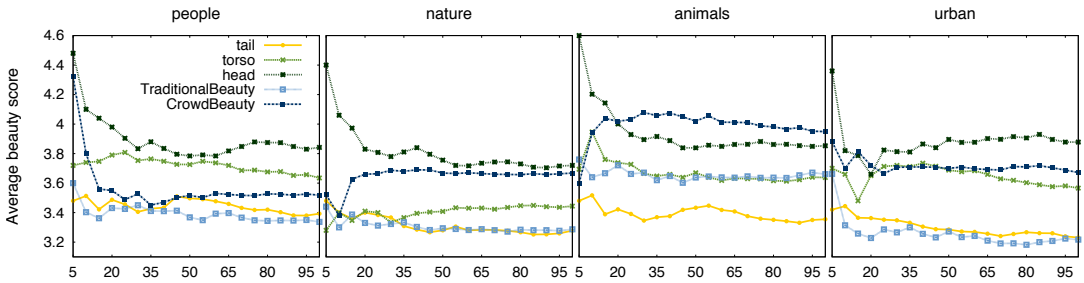

Figure 6: Average beauty of the top n pictures ranked by popularity (in tail, torso, and head buckets) and by the predicted beauty scores.

we used a corpus composed by 200, 200 and 100 images from the tail, torso and head of the popularity distribution respectively, and we added the top 200 images from the *TraditionalBeauty* and *CrowdBeauty* rankings. For consistency, we maintained the same proportion of items per class we used in the previous experiments, but with a smaller sample that focuses only on the top ranked beautiful images.

Figure 5 shows the average beauty score for each category and bucket combination. Consistently across categories, the perceived beauty of the *CrowdBeauty* images is comparable to the most favorited photos. In fact, for *nature* and *animals* we observe an average increment of 0.9% and 1.3% with respect to most popular items and for *urban* and *people* a decrease of 2% and 7%, respectively. With the exception of *people*, the median of the perceived beauty score goes up from 3 to 4 when *CrowdBeauty* is adopted against *TraditionalBeauty*. This behavior confirms how important the training of an aesthetic predictor with a reliable ground truth is for this task.

The study of the average behavior of the beauty predictors does not show what happens if we consider only the head of the rank. For some applications this could be relevant, e.g., recommender systems suggest the top n most relevant items for a user. To this extent it is interesting to evaluate the perceived beauty of the topmost images. Figure 6 shows for each category how the average beauty score varies at cutoffs $n \in [5, 100]$. Highly popular items have a consistent behavior across categories where items at the top of the rank are perceived as very appealing and the quality drops and stabilizes quickly after $n = 20$. In general, after an initial variation, *CrowdBeauty* stabilizes above the tail, torso and *TraditionalBeauty* curves. If *urban* is almost stable for all the cutoffs, *nature* and *animals* start with lower quality items and rapidly jump to higher values. A different case is the *people* category where the top ten images have a very high score and then they drop after n=20.

Some examples of highly ranked images surfaced by our algorithm alongside with the least and most favorited pictures are shown in Table 6.

Table 6: Samples of images from *tail* and *head* popularity buckets, compared to the images surfaced by our approach.

8 Discussion and Conclusions

Applications and future work. The ability to rank by aesthetic appeal images that are nearly indistinguishable in terms of the user feedback by aesthetic value has immediate applications. First, it promotes the *democratization* of photo sharing platforms, creating an opportunity to balance the visibility of popular and beautiful photos with those that are as beautiful but with less social exposure. As a proof-of-concept, we envision a new Flickr *Beauty Explorer* page that surfaces the most beautiful yet unpopular photos of the month to complement the classic Flickr *Explorer* that contains photos with very high social feedback. Our method can be used to bring valuable but unengaged users into the active core of the community by canalizing other people's attention towards them. An extension to this work could be to use the aggregation of photo quality over users to spot hidden *talents* and devise incentive mechanisms to prevent them to churn. Furthermore, our method increases the payoff of the service provider by uncovering valuable content, exploitable for promotion, advertising, mashup, or any other commercial service, that would have been nearly useless otherwise. Also it would be interesting to study the effect of aesthetic reranking on the *head* of the popularity distribution, or on images relevant to a specific query.

Limitations. Our approach comes with a few limitations, mainly introduced by the computer vision method we use.

(a) Animals (b) Urban (c) People

Figure 7: Examples of biases in surfaced pictures.

First, although *machine-tags* have a very high accuracy, they sometimes recognize objects even when they are simply drawn or sketched, and attach semantic tags to non-photographic images, e.g., clipart (see Figure 7c). Non-photographic images have their own aesthetic rules that differ substantially from photographs, and photo aesthetic predictors typically give erroneous predictions on non-photographic images. While in this work we manually removed some non-photographic images from our corpus to allow the model to learn photographic aesthetic rules, an automatic pre-filtering based on non-photographic image detectors would be advisable (Ng, Chang, and Tsui 2007).

Second, despite the high quality of the surfaced photos, some top-ranked *animals* and *nature* images receive lower scores than some lower-ranked ones. This behavior is due to biases in the learning framework: some of the top-rated images for *animals* and *nature* are extremely contrasted pictures (see Figure 7a) thus the model wrongly over-weights the contrast features. Similarly, some of the surfaced *urban* pictures show strong presence of contrast/median filtering, such as the example in Figure 7b.

Last, our method is less effective in surfacing good *people* images. Often highly rated pictures of people show black and white color palette, thus biasing the aesthetic model. From a broader perspective, pictures of people are different in nature from other image types. Faces grasp human attention more than other subjects (Bakhshi, Shamma, and Gilbert 2014): face perception is one of the most developed human skills (Haxby, Hoffman, and Gobbini 2000), and that we have brain sub-networks dedicated to face processing (Freiwald and Tsao 2014). Moreover, when shooting photos of people, photographers need to capture much more than the traits of the mere subject: people come with their emotions, stories, and lifestyles. Portrait photography is indeed a separate branch of traditional photography with dedicated books and compositional techniques (Weiser 1999; Child 2008; Hurter 2007). The traditional compositional features that we use in our framework can only partially capture the essence of the aesthetics of portraits.

Concluding remarks. The popularization of online broadcast communication media, the resulting information overload, and the consequent shrinkage of the attention span online have shaped the Social Web increasingly towards a frantic search for popularity, that many users yearn for. In this rampant race for fame that very few can win, the crowd often cannot see (and sometimes tramples on) some of the valuable gems that itself creates. To fix that in the context of photo sharing systems, we show that it is possible to apply computer vision techniques that spot beautiful images from the immense and often forgotten mass of pictures in the popularity tail. To do that, we show the necessity of using dedicated crowdsourced beauty judgements done by common people on common people's photos, in contrast to corpora of professional photos annotated by professionals. We hope that our work can be a cautionary tale about the importance of targeting content quality instead of popularity, not just limited to multimedia items but in social media at large.

Acknowledgments

R. Schifanella was partially supported by the Yahoo FREP grant. We thank Dr. Judith Redi for her precious help and discussions.

References

Ahmed, M.; Spagna, S.; Huici, F.; and Niccolini, S. 2013. A peek into the future: Predicting the evolution of popularity in user generated content. In *WSDM*.

Aiello, L. M.; Deplano, M.; Schifanella, R.; and Ruffo, G. 2012. People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks. In *ICWSM*.

Baeza-Yates, R. A., and Sáez-Trumper, D. 2013. Online social networks: beyond popularity. In *WWW (Companion Volume)*.

Bakhshi, S.; Shamma, D. A.; and Gilbert, E. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *CHI*.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *WWW*.

Bohn, A.; Buchta, C.; Hornik, K.; and Mair, P. 2014. Making friends and communicating on facebook: Implications for the access to social capital. *Social Networks* 37:29 – 41.

Brodersen, A.; Scellato, S.; and Wattenhofer, M. 2012. Youtube around the world: Geographic popularity of videos. In *WWW*.

Burke, M.; Kraut, R.; and Marlow, C. 2011. Social capital on facebook: Differentiating uses and users. In *CHI*.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S. 2007. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *IMC*.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in Twitter: The million follower fallacy. In *ICWSM*.

Cha, M.; Mislove, A.; and Gummadi, K. P. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*.

Child, J. 2008. *Studio photography: essential skills*. CRC Press.

Chu, W.-T.; Chen, Y.-K.; and Chen, K.-T. 2013. Size does matter: How image size affects aesthetic perception? In *MM*.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*.

Dhar, S.; Ordonez, V.; and Berg, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*.

Figueiredo, F.; Benevenuto, F.; and Almeida, J. M. 2011. The tube over time: Characterizing popularity growth of youtube videos. In *WSDM*.

Freeman, M. 2007. *The Photographer's Eye: Composition and Design for Better Digital Photos*, volume 1. Focal Press.

Freiwald, W. A., and Tsao, D. Y. 2014. Neurons that keep a straight face. *Proceedings of the National Academy of Sciences* 111(22):7894–7895.

Fu, Y.; Hospedales, T.; Xiang, T.; Gong, S.; and Yao, Y. 2014. Interestingness prediction by robust learning to rank. In *ECCV*.

Goel, S.; Broder, A.; Gabrilovich, E.; and Pang, B. 2010. Anatomy of the long tail: Ordinary people with extraordinary tastes. In *WSDM*.

Hagen, M. A., and Jones, R. K. 1978. Cultural effects on pictorial perception: How many words is one picture really worth? In *Perception and Experience*, volume 1 of *Perception and Perceptual Development*. Springer.

Haralick, R. M. 1979. Statistical and structural approaches to texture. *Proceedings of the IEEE* 67(5):786–804.

Haxby, J. V.; Hoffman, E. A.; and Gobbini, M. I. 2000. The distributed human neural system for face perception. *Trends in cognitive sciences* 4(6):223–233.

Hodas, N. O., and Lerman, K. 2012. How visibility and divided attention constrain social contagion. In *PASSAT*.

Hong, L.; Dan, O.; and Davison, B. D. 2011. Predicting popular messages in twitter. In *WWW*.

Hou, X., and Zhang, L. 2007. Saliency detection: A spectral residual approach. In *CVPR*, 1–8. IEEE.

Hsieh, L.-C.; Hsu, W.; and Wang, H.-C. 2014. Investigating and predicting social and visual image interestingness on social media by crowdsourcing. In *ICASSP*.

Hurter, B. 2007. *Portrait Photographer's Handbook*. Amherst Media, Inc.

Isola, P.; Xiao, J.; Torralba, A.; and Oliva, A. 2011. What makes an image memorable? In *CVPR*.

Jiang, Y.-G.; Wang, Y.; Feng, R.; Xue, X.; Zheng, Y.; and Yang, H. 2013. Understanding and predicting interestingness of videos. In *AAAI*.

Karnstedt, M.; Rowe, M.; Chan, J.; Alani, H.; and Hayes, C. 2011. The effect of user features on churn in social networks. In *WebSci*.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *CVPR*.

Khosla, A.; Das Sarma, A.; and Hamid, R. 2014. What makes an image popular? In *WWW*.

Li, C.; Gallagher, A.; Loui, A. C.; and Chen, T. 2010. Aesthetic quality assessment of consumer photos with faces. In *ICIP*.

Luo, Y., and Tang, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *ECCV*.

Luo, W.; Wang, X.; and Tang, X. 2011. Content-based photo quality assessment. In *ICCV*, 2206–2213. IEEE.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *MM*.

Marchesotti, L.; Perronnin, F.; Larlus, D.; and Csurka, G. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 1784–1791. IEEE.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior Research Methods* 44(1):1–23.

McIntosh, A.; Bookstein, F.; Haxby, J. V.; and Grady, C. 1996. Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* 3(3):143–157.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*.

Ng, T.-T.; Chang, S.-F.; and Tsui, M.-P. 2007. Lessons learned from online classification of photo-realistic computer graphics and photographs. In *SAFE*.

Nishiyama, M.; Okabe, T.; Sato, I.; and Sato, Y. 2011. Aesthetic quality classification of photographs based on color harmony. In *CVPR*.

Obrador, P.; Anguera, X.; de Oliveira, R.; and Oliver, N. 2009. The role of tags and image aesthetics in social image search. In *WSM*.

Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2011. In the mood for being influential on twitter. In *SocialCom*.

Redi, M., and Merialdo, B. 2012. Where is the beauty?: Retrieving appealing videoscenes by learning flickr-based graded judgments. In *MM*.

Redi, J., and Povoa, I. 2014. Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd? In *CrowdMM*.

Redi, J. A.; Hoßfeld, T.; Korshunov, P.; Mazza, F.; Povoa, I.; and Keimel, C. 2014a. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *QoMEX*.

Redi, M.; O'Hare, N.; Schifanella, R.; Trevisiol, M.; and Jaimes, A. 2014b. 6 seconds of sound and vision: Creativity in micro-videos. In *CVPR*.

Romero, D. M.; Galuba, W.; Asur, S.; and Huberman, B. A. 2011. Influence and passivity in social media. In *Machine Learning and Knowledge Discovery in Databases*. Springer.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.

San Pedro, J., and Siersdorfer, S. 2009. Ranking and classifying attractiveness of photos in folksonomies. In *WWW*.

Sastry, N. R. 2012. How to tell head from tail in user-generated content corpora. In *ICWSM*.

Siahaan, E.; Redi, J. A.; and Hanjalic, A. 2013. Beauty is in the scale of the beholder: a comparison of methodologies for the subjective assessment of image aesthetic appeal. In *CrowdMM*.

Smith, M., and Giraud-Carrier, C. 2010. Bonding vs. bridging social capital: A case study in twitter. In *SocialCom*.

Steinfield, C.; Ellison, N. B.; and Lampe, C. 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *J. of Applied Developmental Psychology* 29(6):434 – 445.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *SocialCom*.

Totti, L. C.; Costa, F. A.; Avila, S.; Valle, E.; Meira, Jr., W.; and Almeida, V. 2014. The impact of visual attributes on online image diffusion. In *WebSci*.

Vaca Ruiz, C.; Aiello, L. M.; and Jaimes, A. 2014. Modeling dynamics of attention in social media with user efficiency. *EPJ Data Science* 3(1):5.

Weiser, J. 1999. *Phototherapy techniques: Exploring the secrets of personal snapshots and family albums*. PhotoTherapy Centre.

Weng, L.; Flammini, A.; Vespignani, A.; and Menczer, F. 2012. Competition among memes in a world with limited attention. *Scientific Reports* 2.

Wu, O.; Hu, W.; and Gao, J. 2011. Learning to predict the perceived visual quality of photos. In *ICCV*.

Zhong, C.; Shah, S.; Sundaravadivelan, K.; and Sastry, N. 2013. Sharing the loves: Understanding the how and why of online content curation. In *ICWSM*.