Image Popularity on Etsy through Hand-crafted Feature Vectors

Stephen Zakrewsky
Drexel University
sz372@drexel.edu

Kamelia Aryafar
Etsy Inc.
karyafar@etsy.com

Ali Shokoufandeh
Drexel University
ashokouf@cs.drexel.edu

***Abstract*—The online realm has become a driving force in the retail marketplace. E-Commerce websites can provide a level of diversity and uniqueness that is impossible in the world of brick-and-mortar retail. Etsy is an online marketplace[1] for artisans selling unique handcrafted goods, and vintage wares that couldn't be found elsewhere. Etsy caters to the long tail of online retail [1].**

Intuitively, online retail is a visual experience- shoppers have particular styles that they find appealing; often images are used as first order information when making shopping decisions. There are a variety of signals extracted from the images representing those items for sale by shoppers. Amongst these, light, color, texture, scene composition and aesthetics are important cues for visual search and image ranking. In this paper, we introduce a novel dataset for user behaviour prediction. We address the problem of inferring what makes a listing popular from the pixel-level information of listed images on Etsy. We explore the popularity of favorited listings and investigate the entropy of popularity among Etsy users.

I. Introduction

It is difficult to overstate the importance of online retail and e-commerce. According to the website statista[2], 190 million people in the United States alone purchased something online. These purchases covered the spectrum of consumer products- books, clothes, furniture, home goods, electronics, and almost anything else one could imagine. Indeed, the economies of scale provided by online retail enable sales that would otherwise be impossible. If one wants a MiG fighter jet, or a part of the space shuttle, it is available online, provided one knows where to look and that the time is right.

While online retail seems poised to transform the way people shop, the way shoppers connect with their purchases- through touch, through smell, and through sight, aren't possible with online goods. To bridge these gaps, online retailers have emphasized the visual components of their wares, offering detailed descriptions, and, importantly, rich visual representations of what is for sale in the form of images. Because images offer an unparalleled degree of information density to the shopper, they have become a crucial factor in the online shopping decision process.

This paper presents techniques for improving and understanding the online shopping experience at a major e-commerce website, Etsy. Rather than act as a retailer of goods, Etsy is a marketplace, where over a million individual sellers can set up shop, and market their own goods to a shopping community of many millions of people in over 180 countries. Currently in this marketplace, there are over 30 million unique items for sale (listings), and well over 90 million unique images representing these items. With such an expansive image dataset, Etsy is uniquely poised to present users with a rich visual experience.

Importantly, with the proliferation of mobile devices, the text that has dominated conventional information retrieval for the past 15 years has become burden- users are driven by visual experiences rather than through the input and consumption of text. This reliance on visual stimuli has spurred a great interest in new ways to expand textual search into visual search. In doing so, one must understand the behavior of shoppers surrounding images.

The presentation of product listings through images plays an important role on commerce websites. Conventional text-based product search relies heavily on the similarity between the query input by a user and the textual metadata describing what is for sale- listing tags, titles, descriptions, etc. All information added by those users who have listed what is for sale. These search results might include a large number of relevant product listings, all of them containing very similar tags, but with varying levels of image quality and aesthetic appeal. Generally, in the setting of unique handmade or vintage goods, this text-based approach to shopping might yield high recall, but low precision. For a product listing to stand out among query-based search results, high-quality images describing the content of the product listing is a necessity [2], [3].

In this paper we introduce a mechanism for extracting popularity of product listings from the

[1] www.etsy.com
[2] http://www.statista.com/statistics/183755/number-of-us-internet-shoppers-since-2009/

images representing those listings. We then explore the correlation between listings' popularity and user interaction with what is for sale. Because sales are rare in comparison to the number of items available on a large site such as Etsy, we look into an alternative mechanism for interaction, the "favorite." Favorites on Etsy are similar to any number of "like" mechanisms available online, the most familiar of which is Facebook's ubiquitous "thumbs-up." By considering what users explicitly express interest in, we are able to form relationships between user preferences and popularity-based image features.

The remainder of the paper is organized as follows: Section II discusses background on predicting image popularity. Section III describes the features we used to predict populartiy. We examine the performance of popularity features in predicting favorite listings classification in section IV-A and explore the popularity entropy among active users in section IV-B. Finally, we conclude this paper in section V and propose future research directions.

II. Background

Recent advances in multimedia industry have redefined the dynamics of e-commerce. Tens of millions or even billions of product listings and associated listings images are available in a large online market place. These massive datasets require various methods for acquiring, processing, analyzing, and understanding images in order to produce numerical or symbolic information, such as color and texture characteristics that can be used for content based image search and retrieval [4], [5], [6]. Various global image features such as color histograms [7], texture values [8] and shape parameters of easily segmentable regions [9] and localized features such as scale invariant feature transform (SIFT) [10], speeded up robust features (SURF) [11] and histogram of oriented gradients (HOG) [12] are proposed for image content retrieval and search. With content-based image retrieval on the rise, the study of cues that could help in ranking the images of varying levels of popularity is becoming an important problem.

A. What is Popularity?

Early work defined popularity as quality [13] or aesthetics [14] and use data from photography rating websites where users who have interest in photography upload their photos and rate others. Popularity has also been defined as memorability [15], and interestingness [16], [17]. More recent work has directly tackled popularity. In [18], popularity is defined as the number of views on Flickr, and [19] uses favorited listings on Etsy.

B. How to Predict Popularity

Popularity tends to be predicted using SVM classification or regression [14] [18] [20] [21]. Datta et. al. [14] uses a two class SVM classifier with a forward selection algorithm to find good feature sets. By using elastic net to rank feature relevance to aesthetics, and a best first algorithm to find feature sets that minimize the RMSE cross validation error, [21] are able to achieve a 30.1% improvement compared to [20]. A few have explored other machine learning techniques. In [13] a naive Bayes classifier is used, not SVM. Aryafar et. al [19] studied the significance of color in favorited listings on Etsy using logistic regression, perceptron, passive aggressive and margin infused relaxed algorithms.

C. Popularity Features

The features used in popularity prediction model the same qualities professional photographers use such as light, color, rule of thirds, texture, smoothness, blurriness, depth of field, scene composition [13] [14] [20] [21]. Most of these features are unsupervised, but some such as the spacial edge distribution and color distribution features of [13] require all of the labeled training data. Some recent work has looked at semantic object features. [18] used the popular CNN ImageNet to detect the presence of 1000 difference object categories in the image. The presence/absence of these categories is used as the feature.

D. Our Approach

With more than 30 million active listings and over 90 million listing images, Etsy provides a unique visually enticing experience for users. Because images are uploaded by users of the site, representing the myriad items for sale, these images are composed of different items, presented with various lighting conditions, scene geometries and background selections. One of the key components of e-commerce websites is efficient image search and color filtering methods. Presence of occluded backgrounds and highly textured material can hinder the accuracy of color detection algorithms. In our work, we define popularity as listings that have been favorited, clicked on, or purchased, and we show that unsupervised image popularity features are statistically significant when combined with traditional text meta-data features in predicting popularity.

III. Features

A. Simplicity

High quality photos are typically simpler than others. They often have one subject placed deliberately in the frame. Sometimes the background is out of focus to emphasize the subject. Poor quality photographs tend to have cluttered backgrounds and it may be difficult to distinguish the subject of the scene. We used the four measures of simplicity from [13], spatial edge distribution, hue count, contrast and lightness, and blur.

1) Spatial Edge Distribution: Spatial edge distribution measures how spread out sharp edges are in the image. A single subject is expected to have a small distribution while an image with a cluttered background would have a large distribution. Edges are detected by applying a 3x3 Laplacian filter and taking the absolute value. The filter is applied to each RGB channel independently and the final image is computed as the mean across all three channels. The Laplacian image is resized to 100x100 and normalized to sum to 1. Then, the edges are projected onto the x and y axis independently. Let w_x, and w_y be the width of 98% of the projected edges respectively. The image quality feature $f = 1 - \frac{w_x w_y}{100}$ is the percent of area outside the majority of edges. Figure 1 shows the edges detected from two different images and their respective feature value.

2) Hue Count: Professional photographs look more colorful and vibrant, but actually tend to have less distinct hues because cluttered scenes contain many heterogeneous objects. We use a hue count feature by filtering an image in HSV color space such that V is in the range of [0.15, 0.95] and S is greater than 0.2. A 20 bin histogram is computed on the remaining H values. Let m be the maximum value of the histogram and let $N = \{i | H(i) > \alpha m\}$, be the set of bins values greater than αm. The quality feature $f = 20 - ||N||$ is 0 when there are a many different hues and larger as the number of distinct hues in the image goes down. We used $alpha = 0.05$ as in [13].

3) Contrast and Lightness: Brightness is a well known variable that professional photographers are trained to understand and adjust. We use the average brightness feature [13], [20] computed from the L channel of the Lab color space. Contrast is similar, and is the ratio of maximum and minimum pixel intensities. We sum the RGB level histograms, and normalize it to sum to 1. We use the width of the center 98% mass of the histogram [13].

B. Blur

Blurry images are almost always considered to be of poor quality. We use the blur features of [13] and [22]. In [13] blur is modeled as $I_b = G_\sigma * I$ where I_b is the result of convolving a Gaussian filter with an image. The larger the σ the more high frequencies are removed from the image. Assuming the frequency distribution of all I is approximately the same, then the maximum frequency $||C||$ can be estimated as $C = \{(u, v) \mid ||FFT(I_b)|| > \Theta\}$. The feature is $f = ||C|| \sim 1/\sigma$, after normalizing by the image size.

In [22], blur estimation is done based on changes in the edge structures. The blur operation will cause gradual edges to lose sharpness. Assuming that most images have gradual edges that are sharp enough, the blur is measured as the ratio of gradual edges that have lost their sharpness.

C. Rule of Thirds

The rule of thirds is an important composition technique. Thirds lines are the horizontal and vertical lines that divide an image into a 3x3 grid of equal sized cells. The rule of thirds states that subjects placed along these lines are aesthetically more pleasing and more natural than subjects centered in the photograph. In order to segment the subject of the image from the background, we use the Spectral Residual saliency detection algorithm [23]. The feature is a 5x5 map where each cell is the average saliency value [24]. Let w_p be the saliency value of the pixel and $A(W_i)$ is the area of the cell, then the value of each cell is

$$w_i = \frac{\sum_{p \in W_i} w_p}{A(W_i)}. \tag{1}$$

To compute the feature, the image is divided into a 5x5 grid with emphasis on the thirds lines; the horizontal and vertical regions centered on the thirds lines are 1/6 of the image size. Figure 2 shows the saliecy detection with the 5x5 grid overlay, and the thirds map feature for an image.

D. Texture

A smooth image may indicate blur or out-of-focus, and the lack of which may indicate poor film, or too high an ISO setting. In contrast, texture in the scene is an important composition skill of a photographer. Smoothness may indicate the lack of texture. Texture and smoothness are some of the most statically correlated features for quality/popularity [21] and [18]. We use three smoothness/texture features from these.

A three level wavelet transform is applied to the L channel of the Lab color space. We only use the bottom level of the pyramid. The result is squared to indicate power. Let $b = \{HH, HL, LH\}$ be the bottom level of a wavelet transform, the feature is

$$f = \frac{1}{3MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{b} w^b(m, n) \tag{2}$$

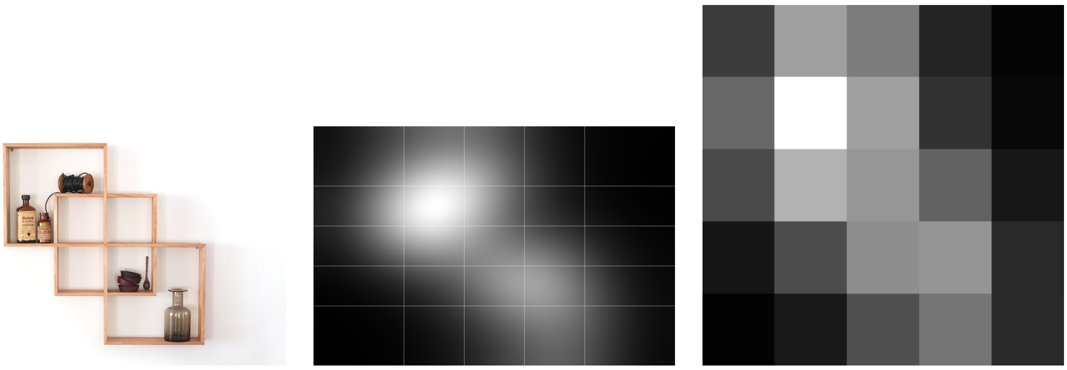
where w is the square of the wavelet value. Because the Laplacian is often used as a pyramid of different scales, another feature

$$f = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} l(m, n) \tag{3}$$

(a)

(b)

Fig. 1: The Laplacian image for computing spacial edge distribution for two images. The feature for figure a. is 0.013 and for b. is 0.30.



(a) (b) (c)

Fig. 2: Example of Rule of Thirds feature. Figure b. shows the SR saliency detection, and c. shows the thirds map feature.

is also used. This time l is the second level from the bottom of a Laplacian pyramid.

Another texture feature is computed using local binary pattern (LBP). Then a pyramid of histograms are computed as in [25]. Figure 3 shows the similarities of LBP features and the three channels of Daubechies db1 wavelet.

E. Depth of Field

Depth of field is the distance between between the nearest and farthest objects that appear in sharp focus. A technique of professional photographers is to use low depth of field to focus on the photographic subject while blurring the background. We used the feature [14] of the ratio of high frequency detail in center regions of the image compared to the entire image. Let w be the bottom level of a wavelet transform, the feature is

$$f = \frac{\sum_{(x,y)} \in M_6 \cup M_7 \cup M_{10} \cup M_{11} w(x,y)}{\sum_{i=1}^{16} \sum_{(x,y) \in M_i} w(x,y)}, \quad (4)$$

where $M_i | 1 \leq i \leq 16$ are the cells of a 4x4 grid. The same feature is also reapplied using the Laplacian pyramid l instead of w [21]. These features only look at the center region of the image. A third feature [21] looks at the spacial distribution of high frequency details. Let l be the bottom layer of a Laplacian pyramid and c_{row}, c_{col} are the center of mass, the feature is

$$f = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} l(m,n) \sqrt{(m - c_{row})^2 + (n - c_{col})^2}. \quad (5)$$

Figure 4 visualizes how these features are computed for an image.

F. Experimental

Maximally Stable Extremal Regions (MSER) [26] can be used to detect text because characters are typically single solid colors with sharp edges that standout from the background [27]. Additionally, texture patterns are also often detected by MSER, like bricks on a wall. We used the experimental feature the count of the number of MSER regions. We would like to continue this experiment into other features based on text in images.

IV. User Behaviour and Colors

In this section we present the unique dataset for quantitative evaluation of colors impact in user behaviour on Etsy. Once an item is listed on Etsy, the users can favorite a listing which allows them to bring all the items they like in one place. We first examine the listings dominant colors to predict if a listings is likely to be favorited or not. Then we explore the entropy of dominant colors among users favorites to indicate the color variations among favorited listings. Two datasets are used for classification of favorited listings and color entropy experiments. The classification dataset consists of 2.73 million unique listings that have been created within the last month on Etsy. The listings images are tagged with the top three dominant colors and labeled as positive if they have been favorited by a user in that period. To collect the entropy dataset, a set of 11235 active users with more than 20 and less than 2000 favorited listings within the past six months are selected. The entropy dataset then consists of 2.32 million unique listings that candidate users have favorited over the last 6 months on Etsy. These listings images are also tagged with top three dominant colors. These two datasets contain more than 5.05 million listing images and dominant color tags and are available through Etsy's API[3].

Feature	Dimension
'Ke06-qa': spatial edge distribution	1
'Ke06-qh': hue count	1
'Ke06-qf': blur	1
'Ke06-tong': blur tong etal	1
'Ke06-qct': contrast	1
'Ke06-qb': brightness	1
'-mser count': mser count	1
'Mai11-thirds map': thirds map	25
'Wang15-f1': avg lightness	1
'Wang15-f14': wavelet smoothness,	1
'Wang15-f18': laplacian smoothness	1
'Wang15-f21': wavelet low dof	1
'Wang15-f22': laplacian low dof	1
'Wang15-f26': laplacian low dof swd	1
'Khosla14-texture': texture	5120

TABLE I: Feature Dimensions

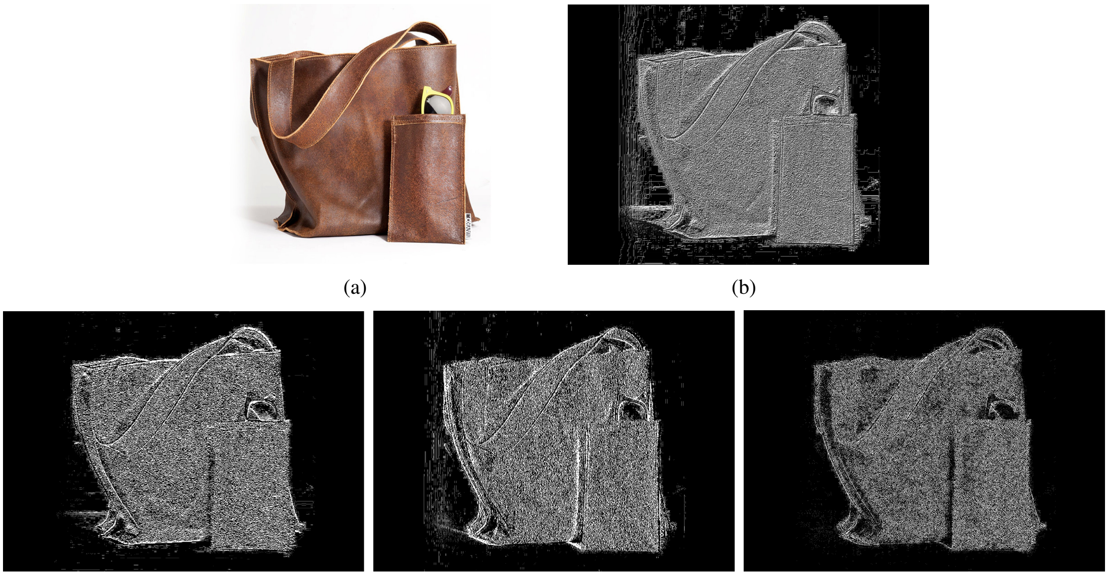
A. Classification

Classification method	Average accuracy rate	AUC
logistic regression	0.5512	0.5694
perceptron	**0.5600**	**0.5906**
passive aggressive	0.5329	0.5240
MIRA	0.5232	0.5240

TABLE II: Average classification accuracy rate and AUC are reported for favorite listing classification using text and color features.

Once the classification dataset has been tagged with top three dominant colors, we extract textual information from the listings. These textual features consist of the tokenized listings titles unigrams and bigrams and tokenized listings tags unigrams. We then represent each listing with a feature vector including textual features and color unigrams. A binary classification is then performed to predict if the test listings are favorited by users. We report the average

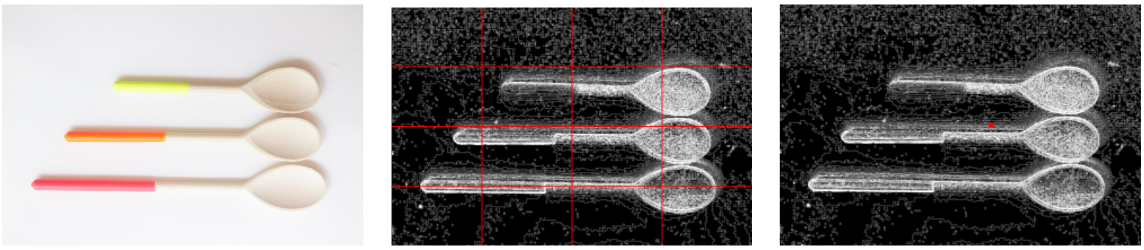
[3] www.etsy.com/developers



(a) (b)

(c)

Fig. 3: Smoothness and texture features. Figure b. shows Local Binary Pattern (LBP) feature image, and c. shows the 3 channels of the DB1 wavelet transform.



(a) (b) (c)

Fig. 4: Figure b. shows the Low Depth of Field features in the center grid region for the Laplacian image. Figure c. shows the same image with its center of mass.

classification accuracy rate and the area under the curve (AUC) with four different classifiers. Logistic regression, passive-agressive classifier, perceptron and margin infused relaxed algorithm (MIRA) are used as the learning models. Table II shows the results of four rounds of 5-fold classification on this dataset. The textual information and color unigrams do not indicate a strong improvement in favoriting behaviour prediction. It will be interesting to observe the effects of image quality, memorability, aesthetics and interestingness for the similar problem on this unique dataset.

B. *Entropy Estimation on Selected Users*

In this experiment we measure the entropy of dominant colors distribution among 11235 active users with more than 20 and less than 2000 favorited listings within the past six months. These users have favorited 2.32 million unique listings. The listings images are first tagged with top three dominant colors and the entropy of color distribution is estimated using the histogram approach. Figure **??** shows the number of users in each entropy interval. As we can observe in figure **??**, most active users have high color distribution entropy in their favorited listings. There are however, some users with low color distribution entropy which highlights a strong tendency to favorite specific colors on Etsy. Figure **??** shows a set of favorites for a user with the lowest color distribution entropy while figure **??** illustrates selected favorited listings by highest color entropy user.

V. CONCLUSION

This works represents a initial study on understanding how images, specifically color-based image features, can be used to represent items in an e-commerce setting, thereby providing better user understanding and a better overall shopping experience. To facilitate this understanding, this work proposed an empirical method to estimate the dominant colors of images representing product listings on Etsy, using object localization.

We used this dominant colors to filter listings in a conventional text-based e-commerce search, according to user input. Moreover we explored the color distribution among candidate users favorited listings. We also examined the impact of color unigrams on users favoriting behaviour. This work represents the tip of the iceberg of understanding how visual cues influence user action in an e-commerce setting.

While color is no doubt important and an influencing factor while shopping, future work involves incorporating a deeper visual understanding of listing images. Image composition, texture and aesthetics all influence the emotional reaction that users have when seeing an item for the first time. Future work seeks to incorporate a richer understanding of images, using this improved representation to better model user preference, and provide a more effective e-commerce experience.

REFERENCES

[1] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[2] Y. J. Wang, M. S. Minor, and J. Wei, "Aesthetics and the online shopping environment: Understanding consumer responses," *Journal of Retailing*, vol. 87, no. 1, pp. 46–58, 2011.

[3] P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver, "The role of tags and image aesthetics in social image search," in *Proceedings of the first SIGMM workshop on Social media*. ACM, 2009, pp. 65–72.

[4] A. Bhardwaj, A. D. Sarma, W. Di, R. Hamid, R. Piramuthu, and N. Sundaresan, "Palette power: Enabling visual search through colors."

[5] C. Zhou, Y. Li, B. Geng, and C. Xu, "From document to image: Learning a scalable ranking model for content based image retrieval," in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*. IEEE, 2012, pp. 558–563.

[6] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *Journal of intelligent information systems*, vol. 3, no. 3-4, pp. 231–262, 1994.

[7] A. Rao, R. K. Srihari, and Z. Zhang, "Spatial color histograms for content-based image retrieval," in *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*. IEEE, 1999, pp. 183–186.

[8] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[9] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "Qbic project: querying images by content, using color, texture, and shape," in *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1993, pp. 173–187.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[13] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 419–426.

[14] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 288–301.

[15] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 145–152.

[16] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1657–1664.

[17] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," 2013.

[18] A. Khosla, A. Das Sarma, and R. Hamid, "What makes an image popular?" in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 867–876.

[19] K. Aryafar, C. Lynch, and J. Attenberg, "Exploring user behaviour on etsy through dominant colors," in *2014 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1437–1442.

[20] M. Chen and J. Allebach, "Aesthetic quality inference for online fashion shopping," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 902 703–902 703.

[21] J. Wang and J. Allebach, "Automatic assessment of online fashion shopping photo aesthetic quality," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2915–2919.

[22] H. Tong, M. Li, H. Zhang, and C. Zhang, "Blur detection for digital images using wavelet transform," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 17–20.

[23] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[24] L. Mai, H. Le, Y. Niu, and F. Liu, "Rule of thirds detection from photograph," in *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 91–96.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[26] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[27] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2609–2612.