Item Popularity Prediction in E-commerce Using Image Quality Feature Vectors

Stephen Zakrewsky
Drexel University
sz372@drexel.edu

Kamelia Aryafar
Etsy
karyafar@etsy.com

Ali Shokoufandeh
Drexel University
ashokouf@cs.drexel.edu

***Abstract*—Online retail is a visual experience- Shoppers often use images as first order information to decide if an item matches their personal style. Image characteristics such as color, simplicity, scene composition, texture, style, aesthetics and overall quality play a crucial role in making a purchase decision, clicking on or liking a product listing. In this paper we use a set of image features that indicate quality to predict product listing popularity on a major e-commerce website, Etsy [1]. We first define listing popularity through search clicks, favoriting and purchase activity. Next, we infer listing quality from the pixel-level information of listed images as quality features. We then compare our findings to text-only models for popularity prediction. Our initial results indicate that a combined image and text modeling of product listings outperforms text-only models in popularity prediction.**

I. Introduction

The informative presentation of product listings through text and images is the foundation of modern e-commerce. Shoppers often have a specific style or visual preference for many of the available items such as jewelry, clothing, home decor, etc. Images provide the first order information for product listings. Users often use images in combination with other data modalities such as textual description, price, ratings and etc. to decide if an item is a suitable match for what they need and have in mind. The selection of proper high quality images is then an important step in listing a successful product. In this paper we examine the role of image quality in listing popularity on a major e-commerce website, Etsy [1].

Etsy is an online marketplace 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], [2]. With more than one million sellers, 35 million unique product listings and nearly a hundred million images, Etsy is uniquely positioned to answer some interesting questions about the role of images as a rich visual experience in e-commerce settings. Each Etsy listing is composed of text information such as title, tags, item description, shop and seller name and complementary images. For a product listing to stand out, high-quality images describing the content of the product listing is a necessity [3], [4]. Figure 1 illustrates some Etsy images with different scene composition, lighting and image quality as featured on the website.

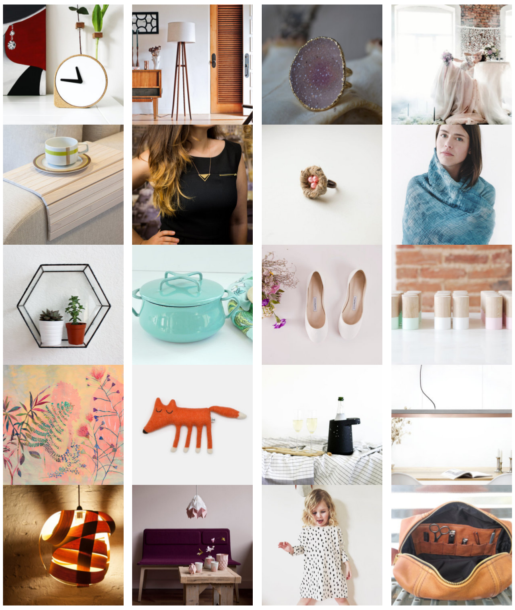


Fig. 1: Sample Etsy listing images are shown with different lighting, scene composition, and quality.

Early work in the literature has defined image popularity as quality [5] or aesthetics [6] 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 [7], and interestingness [8], [9]. More recent work has directly tackled popularity. In [10], popularity is defined as the number of views on Flickr, and [2] uses favorited listings on Etsy.

In this paper we introduce a mechanism for product listings popularity prediction from the images representing those listings. We then explore the correlation between image quality and user interaction with what is for sale. Because sales are rare in comparison to the number of items available on a

[1] www.etsy.com

large site such as Etsy, we look into a combination of mechanisms for interaction, including the number of favorites, purchases and clicks on items to define item popularity. Favorites indicate an interest in an item and are similar to liking mechanisms on other websites such as thumbs-up on Facebook.

Popularity tends to be predicted using typical classifiers such as SVMs or regression [6] [10] [11] [12]. Datta et. al. [6] uses a two class SVM classifier with a forward selection algorithm to find suitable feature vectors indicating popularity. 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, [12] are able to achieve a 30.1% improvement compared to [11]. A few have explored other machine learning techniques. In [5] a naive Bayes classifier is used and Aryafar et. al [2] studied the significance of color in favorited listings on Etsy using logistic regression, perceptron, passive aggressive and margin infused relaxed algorithms.

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, and scene composition [5] [6] [11] [12]. Most of these features are unsupervised, but some such as the spatial edge distribution and color distribution features of [5] require all of the labeled training data. Some recent work has looked at semantic object features. [10] 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. In this paper we propose a combination of simplicity, blur, depth of field, rule of thirds and texture features as the image quality representation. We also combine the image representation with text features as a multimodal embedding of items for sale. State-of-the-art studies have often shown that multimodal embeddings of items can outperform single modality representations for multiple prediction, ranking and classification problems [13] [14] [15].

The remainder of the paper is organized as follows: Section II describes the image quality feature vectors. We examine the performance of image quality features in predicting listing popularity in section III. Finally, we conclude this paper in section IV and propose future research directions.

II. Features

The quality features extracted from images are composed of a set of hand-crafted features including simplicity, blur, depth of field, rule of thirds, experimental and texture features. In this section, we explain the details of each subset of features. The implementation of this features is made publicly available [2]. The final image quality feature vector is a concatenation of these features. Table I shows the dimensionality of each feature. The dimensionality of the final quality feature vector (image representation) is the sum of all these features.

A. Simplicity

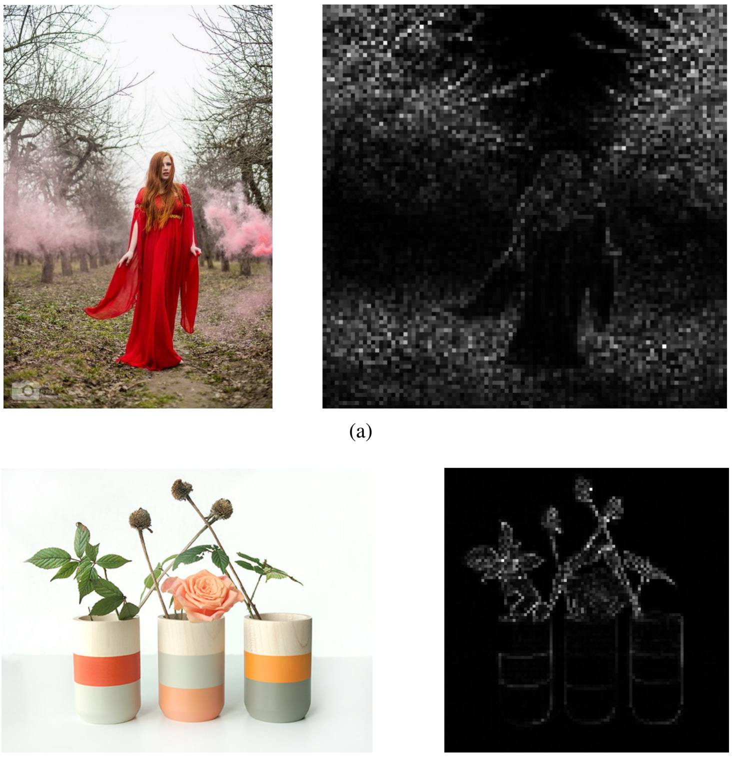
High quality photos are typically simpler than others. They often have one subject placed deliberately in the frame. Sometimes the background is out of focus to emphasize the subject. Poor quality photographs tend to have cluttered backgrounds and it may be difficult to distinguish the subject of the scene. We used the four measures of simplicity from [5], 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 3×3 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 100×100 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 2 shows the edges detected from two different images and their respective feature values.

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 the 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 with values greater than αm. The quality feature $f = 20 - ||N||$ is 0 when there are a many different hues and grows larger as the number of distinct hues in the image goes down. We used $alpha = 0.05$ as shown in the literature [5].

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 [5], [11] computed from the L channel of the Lab color space. Contrast is similar, and is the ratio of maximum and minimum pixel

[2] We make our feature extraction pipeline for image quality features available at:
https://github.com/szakrewsky/quality-feature-extraction



(a)

(b)

Fig. 2: The Laplacian image for computing spatial edge distribution for two images is illustrated. The value of the feature for figure a. is 0.013 and for b. is 0.30.

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 [5].

B. Blur

Blurry images are almost always considered to be of poor quality. We use the common blur features in the literature [5] [16]. In [5] 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.

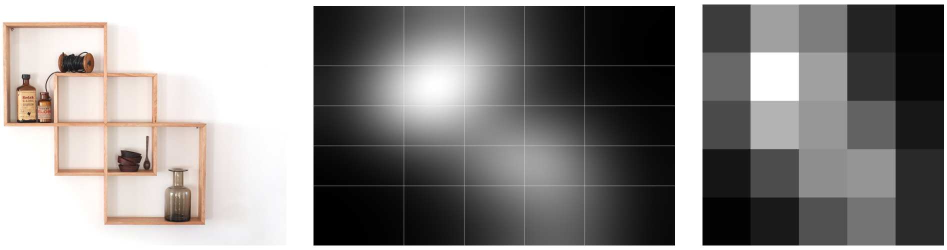
In [16], blur estimation is done based on changes in the edge structures. The blur operation will cause gradual edges to lose sharpness. Assuming that most images have gradual edges that are sharp enough, the blur is measured as the ratio of gradual edges that have lost their sharpness.

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 3×3 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 [17]. The feature is a 5×5 map where each cell is the average saliency value [18]. 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 5×5 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 3 shows the saliency detection with the 5×5 grid overlay, and the thirds map feature for an image.



(a)

(b)

(c)

Fig. 3: Example of Rule of Thirds feature. Figure b. shows the SR saliency detection, and c. shows the thirds map feature.

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 [12] [10]. 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 extracted feature is then

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

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

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

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 [19]. Figure 4 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 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 [6] 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 can be describes as:

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

where $M_i | 1 \leq i \leq 16$ are the cells of a 4×4 grid. The same feature is also reapplied using the Laplacian pyramid l instead of w [12]. These features only look at the center region of the image. A third feature [12] looks at the spatial 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 obtained as:

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

Figure 5 visualizes how these features are computed for a sample image.

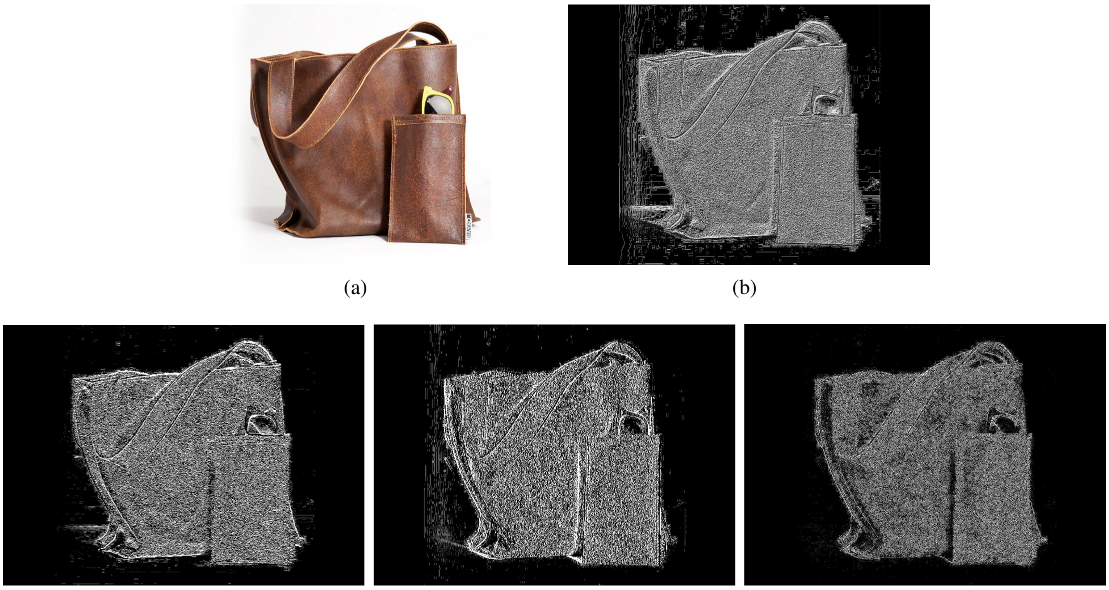
F. Experimental

Maximally Stable Extremal Regions (MSER) [20] can be used to detect text because characters are typically single solid colors with sharp edges that standout from the background [21]. Additionally, texture patterns are also often detected by MSER, like bricks on a wall. In this paper, we used the count of the number of MSER regions as the experimental feature. In the future, we would like to continue this experiment into other features based on text in images.

III. Popularity Prediction

We collect a set of images from Etsy through Etsy's API[3] for popularity prediction. Our dataset consists of $50,000$ Etsy listing images. Each Etsy listing has at least one photo and can have up to five photos to show different angles and details. In our experiments we only extract the first (main)

[3] www.etsy.com/developers



(a)

(b)

(c)

Fig. 4: Smoothness and texture features are illustrated. Figure b. shows Local Binary Pattern (LBP) feature image, and c. shows the 3 channels of the DB1 wavelet transform on the sample image.

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: Image quality feature dimensions are shown by feature.

TABLE II: Lift in accuracy rate using a logistic regression, relative to text-only baseline (%), on the sample dataset is shown in image-only and multi-modal settings.

Modality	Image	Image+Text (MM)
Relative lift in AUC	$+1.07\%$	$+\mathbf{3.45}\%$

listing image which shows up in search results and is featured as the main image on the listing page. We denote the number of favorites for each listing, L with main image I as $F(L_I)$, the number of purchases with $P(L_I)$ and number of clicks with $C(L_I)$. We associate each listing image with it's popularity score as :

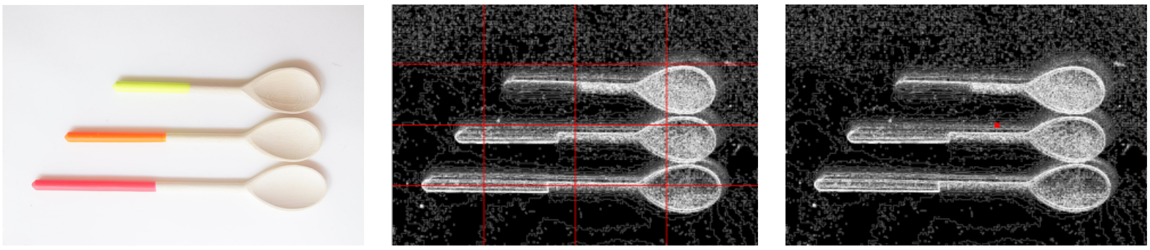
$$Popularity(L_I) = \sum F(L_I) + C(L_I) + P(L_I).$$

We extract the quality feature vectors as described in Section II for each listing image and denote that with $q(L_I)$ for listing L and image I. Table I shows the dimensionality of each feature that is used to build the quality feature vector. Once the dataset has been tagged with these quality features, we extract textual information from the listing as $t(L_I)$. These textual features consist of the tokenized listings titles unigrams and bigrams and tokenized listings tags unigrams and serve as the single modality listing representation. The multimodal feature vector representation, $MM(L_I)$ is obtained by concatenating quality and textual features as a single feature vector, i.e., $MM(L_I) = \langle q(L_I), t(L_I) \rangle$.

We then use a logistic regression against popularity scores, $Popularity(L_I)$ and report the accuracy lift using images and multimodal feature vectors relative to the baseline text-only model. Table II shows these results. We can observe that the quality features in combination with textual features can increase the prediction accuracy on the collected dataset.

IV. Conclusion

This works presents an initial study on understanding how image quality can impact the popularity



(a)

(b)

(c)

Fig. 5: 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.

of items in e-commerce settings, thereby providing better user understanding and a better overall shopping experience. To facilitate this understanding, this work proposed an empirical method to estimate the image quality features representing product listings on Etsy. These feature vectors were combined with traditional textual features to serve as the multimodal item representation. We compared the efficiency of single modality (text-only and image-only) features to multimodal feature vectors in popularity prediction. Our initial results indicate that quality features in combination with text information can increase the prediction accuracy for a sample dataset.

References

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

[2] 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. 1, 2

[3] 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. 1

[4] 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. 1

[5] 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. 1, 2, 3

[6] 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. 1, 2, 4

[7] 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. 1

[8] 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. 1

[9] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," 2013. 1

[10] 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. 1, 2, 4

[11] 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. 2

[12] 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. 2, 4

[13] C. Lynch, K. Aryafar, and J. Attenberg, "Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank," *arXiv preprint arXiv:1511.06746*, 2015. 2

[14] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 2019–2032, 2014. 2

[15] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *Cybernetics, IEEE Transactions on*, vol. 45, no. 4, pp. 767–779, 2015. 2

[16] 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. 3

[17] 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. 3

[18] 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. 3

[19] 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. 4

[20] 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. 4

[21] 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. 4