A COMPARATIVE STUDY OF COMPUTATIONAL AESTHETICS

Dogancan Temel and Ghassan AlRegib

Center for Signal and Information Processing (CSIP)
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, 30332-0250 USA
{cantemel,alregib}@gatech.edu

ABSTRACT

Objective metrics model image quality by quantifying image degradations or estimating perceived image quality. However, image quality metrics do not model what makes an image more appealing or beautiful. In order to quantify the aesthetics of an image, we need to take it one step further and model the perception of aesthetics. In this paper, we examine computational aesthetics models that use hand-crafted, generic and hybrid descriptors. We show that generic descriptors can perform as well as state of the art hand-crafted aesthetics models that use global features. However, neither generic nor hand-crafted features is sufficient to model aesthetics when we only use global features without considering spatial composition or distribution. We also follow a visual dictionary approach similar to state of the art methods and show that it performs poorly without the spatial pyramid step.

***Index Terms*—** Computational Aesthetics, Image Quality, Photography, Color

1. INTRODUCTION

Everyday, we are exposed to various images and video thanks to Facebook, Flickr, Youtube, Instagram and others. The content in these websites or applications are provided by the users. While uploading the multimedia content, files need to satisfy basic constraints such as format, size and resolution. However, these social media platforms do not assess the quality of multimedia content. Subjective quality evaluation is by far the best way to asses the quality of multimedia. However, it requires extensive amount of time and labor. Therefore, objective quality metrics are used to estimate subjective quality.

Most of the quality assessment methods estimate the quality by calculating the degradations in the images and videos. Fidelity-based metrics calculate the accuracy of the processed content with respect to the original content whereas structural metrics model the perceived quality of visual data by considering the Human Visual System (HVS). However, fidelity and structure-based metrics are not sufficient to estimate the quality of experience for the end user. We claim in our work that we also need to consider the aesthetics within images and videos. Hence, in our work, we develop image quality measures that incorporate aesthetics as well as other structure-based and statistical models.

The authors in [1] proposed a computational approach to study aesthetics in photographic images. They studied aesthetics as a machine learning problem by extracting low-level features based on rules of thumbs in photography, common intuition and rating patterns. In [2], authors designed a colorfulness index to asses the quality rather than fidelity. Colorfulness index quantifies the colorfulness in natural images to estimate perceived quality by using the color distribution in the CIELab color space. In addition to the colorfulness feature, brightness, contrast, saturation and saliency were also used to asses the beauty rating of videos in [3]. In [4], the authors used composition, content and sky-illumination as high level attributes to predict aesthetics and interestingness.

Instead of focusing on the entire image, the authors in [5] extracted subject regions using blind motion deblurring [6] to detect areas that draw the most attention of HVS. The authors in [7] used hue and scene composition features as global features whereas dark channel, face region and complexity as regional features. Sharpness, colorfulness, luminance, color harmony and blockiness were used in [8] as low-level features to model visual aesthetic appeal. Composition-specific features such as relative foreground position and visual weight ratio were used in [9] to asses photo quality and perform semi-automatic enhancement based on visual aesthetics. Instead of hand-crafting features that highly correlate with photographic practices and techniques, the authors in [10] used generic image descriptors such as GIST [11] and BOV ([12], [13]) to asses aesthetics of images.

It is not straightforward to describe aesthetics. Therefore, we need to focus on the judgement of subjects. The authors in [5] generated a video database by selecting videos from YouTube. The database contains 4000 high quality professional movie clips and 4000 low quality amateurish clips. The quality of the videos was assessed per frame basis and the average was computed to asses the quality of videos. The authors also used MSN Live Search to search for images and volunteers ranked the images on a scale between 1 and 5. A photography database with peer-rated aesthetics scores ranging from 1 to 7 was provided in [14]. Similarly, [15] provided a large photo database with peer ratings based on quality ranging from 1 to 10. Authors in [16] randomly collected large samples from [17], [15] , [14] and [18] that were annotated with aesthetics, quality, liking and emotion scores. Around 1 million images crawled by Flickr with textual tags, aesthetics annotations, and EXIF meta-data were provided in [19]. A large set of standardized, emotionally-evocative color photographs were provided with a wide range of semantic categories in [20]. Authors in [21] generated a large scale database with score distributions, semantic and style labels and rich annotations including aesthetics.

In this work, we focus on the binary classification of the images based on aesthetics quality. We examine the descriptors used in the aesthetics image quality literature as well as other commonly used image descriptors. In section 2.1, we introduce state of the art computational aesthetics descriptors. Geometric descriptors are discussed in section 2.2 and color descriptors are explained in 2.3. We examine hybrid descriptors in section 2.4 and visual dictionary approach in section 2.5. We compare the descriptors in section 2.6 and conclude our discussion in section 3.

2. AESTHETICS DESCRIPTORS

We can measure the success of the computational models by how good they can estimate the average subjective scores. From the image sets referred in section 1, we use the CUHK database collected by the authors in [22]. The images in the database are obtained from the photo contest website DPChallenge [15] along with the user ratings. From the obtained 60,000 images, the top 10% images are selected as good and the bottom 10% images are selected as bad images. Half of the image set is randomly selected to be used for training with labels and the other half is used for the classification tests. In the following sections, we briefly introduce the descriptors and examine their classification performances.

2.1. State of the art descriptors

Ke at al. [22] designed aesthetics features using spatial distribution of edges, color distribution, hue count, blur, contrast and brightness. Datta et al. [1] used exposure of light, colorfulness, saturation, hue, rule of thirds, familiarity measure, wavelet based texture feature, size and aspect ratio, region composition, depth of field and shape convexity to asses image aesthetics. Tong et al. [23] implemented a black-box approach by generating a set of low-level features and fusing these features using learning algorithms. In addition to extracting global features, Luo and Tang [5] extracted subject regions to obtain local features of foreground and background. They calculated clarity contrast, lighting, simplicity, composition geometry and color harmony to model the aesthetics. Marchesotti et al. [10] used generic features instead of hand-crafted features to perform aesthetics-based classification. SIFT and color features were extracted and a global model was generated using Gaussian Mixture Model (GMM). Bag of Words (BoW) and Fisher Vector (FV) were used to obtain the statistics of feature distributions to train a Support Vector Machine (SVM) classifier.

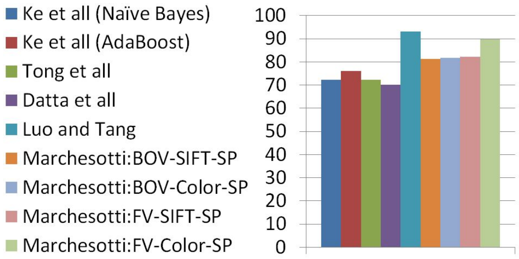

Fig. 1. Classification accuracy of state of the art methods

We summarize the classification accuracy of state of the art methods in Table 1. Hand-crafted global features can reach to a classification accuracy up to 76% with Ke et al. [22]. Local features proposed by Luo and Tang [5] leads to an accuracy up to 93%. Generic features used by Marchesotti et al. [10] result in an accuracy between 81.4% and 89.9%. In the rest of the simulations, we use L1 soft-margin SVM classifier. We experimented other classifiers and observed that classification accuracy did not change significantly when we had a large set of training and test images.

2.2. Geometric Descriptors

GIST is a holistic representation of an image to model the shape of the scene [11]. Scene representation is classified with respect to naturalness, openness, roughness, expansion and ruggerdness. GIST uses the local and global energy spectrum to quantify the introduced metrics in the scene representation. We use three different configurations of the GIST features by varying the number of orientations per scale and number of blocks. GIST(16) and GIST(32) correspond to the configuration where we use an unlocalized energy spectra by setting the number of block to 1. GIST(512) corresponds to the case where the number of blocks is set to 4. Orientations per scales are all set to 8 for GIST(32) and GIST(512) where orientation scales are set to 4 for GIST(16). SIFT divides the image into 4x4 grids of cells and calculates histogram of image gradient directions as explained in [24]. GMM approximates the distributions by weighted sum of Gaussian models. In addition to being used as a feature, it is also used to generate the visual dictionary in section 2.5. Maximally stable extremal regions (MSER) thresholds the image in the intensity channel. Threshold is swept from black to white to detect the connected areas that are unchanged over a large set of thresholds [25]. Difference of Gaussians (DOG) convolves the original image with Gaussian kernels and subtracts the blurred images from each other. Difference of blurred images contains band-pass details that are used as image descriptors. We also detect corners and blobs to represent images using Hessian, Harris and Laplace operators. Implementation of

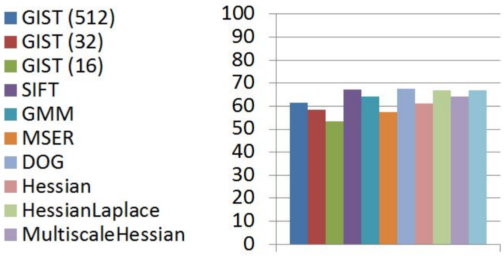

Fig. 2. Classification accuracy of Geometric Descriptors

most of the geometric descriptors are provided with VLFeat package which is an open and portable library of computer vision algorithms [26].

Classification results for geometric descriptors are given in Figure 2. In here, we calculate the geometric descriptors by computing the average of each feature dimension over the whole image. We obtain the most accurate classification with 67.7% in DOG, 67.2% in SIFT and 67.1% in HessianLaplace. Generic geometric descriptors inherently contain information related to the spatial complexity and composition. Some of the geometric descriptors perform better than the others but none of them is sufficient for aesthetics classification since they do not focus on the basic dimensions of the aesthetics.

2.3. Color Descriptors

Color descriptors are designed according to four main constraints: photometric robustness, geometric robustness, photometric stability and generality [27]. In order to satisfy these constraints, color descriptors should be invariant to shadow, shading, light source configuration, view point, orientation and image quality. However, it is not possible to design descriptors that can satisfy all the constrains. Since image aesthetics is also influenced by these factors, we can use color descriptors as aesthetics metrics.

Color naming is introduced in [28] as an image descriptor, which calculates the color distributions similar to the bag of words approach. Relative locations of pixels do not effect the distribution since this approach only focuses on the color distribution of the pixels in the region of interest. It is originally used to assign linguistic color labels to image pixels and the main objective is to predict the color category that humans would perceive given a color measurement. In practice, color naming descriptor is a 11-D vector where each dimension corresponds to the distribution of main colors that can be sorted as follows: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. We also use the color naming method described in [29]. Authors use a fuzzy k-means algorithm to obtain color labels from Munsell book of color. Color descriptor consists of member functions which map color elements to $[0, 1]$ interval and the abbreviation JOSA is used to represent this descriptor in the results .

In addition to the distribution of colors, we can also consider the relative locations of the color pixels. We use discriminative color descriptor introduced in [27] to cluster the color pixels in compact representations. Color pixels are clustered by maximizing the discriminative power using an information theocratical approach. We also use the color descriptors in the opponent color space as described in [30]. Color descriptors are designed in a similar way to SIFT by calculating the histogram of gradients in hue. The hue color descriptor performs poorly when the saturation is low. In case of low saturation, we can use opponent color angle [30].

We use the default version of the color descriptors defined in [31]. Color descriptor matrix is composed of three rows where each row corresponds to the metric calculated over three constant regions. In addition, we take the average of the metric over three regions and use it as an additional descriptor which is shown with a suffix $(1x11)$. In case of discriminative color, we also calculate the color descriptors for a color dictionary size of 25 and 50.

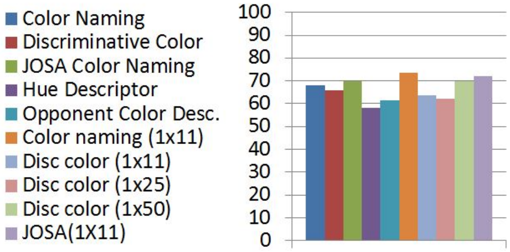

Fig. 3. Classification accuracy of Color Descriptors

Color naming (1x11) results in the highest classification accuracy with 73.6% followed by JOSA(1x11) with 72.1%. When we compare with the generic descriptors, color descriptors perform better. In the literature, variations of color distribution and harmony are commonly used as hand-crafted features to model aesthetics as it is mentioned in section 1.

2.4. Hybrid Descriptors

Color, geometric and aesthetics descriptors analyze the images in different aspects and they can be complementary to each other for classification. However, they can also contradict with each other in terms of classification decisions. We combine some of the descriptors defined in the previous parts to obtain a hybrid descriptor. Classification accuracies of the hybrid descriptors are shown in Figure 4.

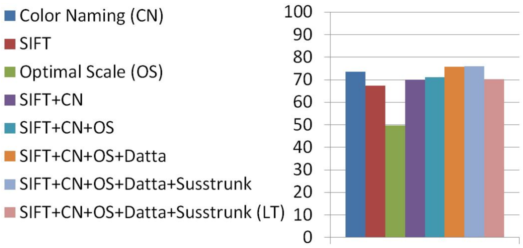

Fig. 4. Classification accuracy of Hybrid Descriptors

We use Color naming (1x11) as the color descriptor which is abbreviated as CN and SIFT is used as the geometric de-

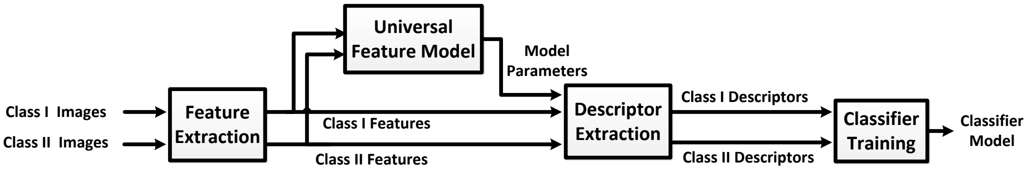

Fig. 5. Training pipeline

scriptor. When CN and SIFT are combined together, classification accuracy gets lower than the accuracy of CN but higher than the accuracy of SIFT with 70.1%. Optimal scale (OS) is a feature introduced to quantify the amount of low frequency content in the image. Classification ratio increases by 1% when OS is added to the descriptor. We add exposure of light, hue, rule of thirds, wavelet-based texture and size and aspect ratio features introduced by Datta et al. [1] to our descriptor. Classification accuracy increases up to 75.7%. Finally, we add brightness, saturation, average color and colorfulness features proposed by Susstrunk et al. [2] and [3]. Our hybrid descriptor has a classification accuracy of 75.9%. Instead of training with the whole set, we train with the first 100 good and bad images to experiment the effect of training set size. Classification accuracy drops from 75.9% to 70.2%.

2.5. Visual Dictionary Approach

In the previous sections, we extract features and combine them in a vector to form a descriptor. However, we do not use the distribution of the features. In contrast, in this section, we generate a visual dictionary to consider the distributional statistics of the features. The training pipeline including visual dictionary generation is shown in Figure 5.

We extract features from two different classes of images and these features are fed to a universal feature model to obtain a visual dictionary. Training features and universal model parameters are fed to a descriptor extraction module to obtain class I and class II descriptors. We train the classifier with the labeled descriptors to obtain a classifier model. We use SIFT, DOG, Color naming (1x11) and JOSA because they have the highest classification accuracy based on our experiments from previous sections. In addition to the default SIFT, we perform singular value decomposition to keep the first N eigenvalues and remove the rest to reconstruct the feature vector. We experimented with different N values and 30 produced the highest classification accuracy. Gaussian Mixture model (GMM) is used as the universal feature model and Fisher Vector is preferred as the descriptor. In our simulations, we vary the number of Gaussians in the GMM and the training set size. For each feature, we perform simulations at least for 14 different configurations and up to 22. Configurations that lead to highest classification accuracy are given in Table 1.

Table 1. Classification accuracy for the dictionary approach

Features Types	Number of Gaussians	Training Set Size	Classification Accuracy (%)
SIFT(SVD)	200	100	75.5
SIFT	200	100	74.0
CN	5	6000	72.6
DoG	2	6000	69.7
JOSA	5	6000	67.6

2.6. Descriptor Comparison

We observe that state of the art descriptors using local information leads to classification ratios up to 90%. Luo and Tong [5] used hand-crafted features and Marchesotti et al. [10] used generic features to obtain high classification accuracy. As it was claimed in [10], generic features can perform as good as hand-crafted features. However, examining local features in addition to global features have a more significant effect on the classification results than feature selection. The maximum classification ratio we obtain using geometric descriptors is 67.7% and it is 73.6% using color descriptors. Hybrid descriptors lead to 75.9% and the dictionary approach leads to 75.5% at most. Basic generic features such as color and geometric can perform as well as the state of the art global computational approaches. However, they do not perform as well as the ones that take spatial characteristics into account by using subject region extraction or spatial pyramid. The main disadvantage of the regional methods comes from the time and memory complexity. Subject region extraction used in [5] is an exhaustive approach that requires significant amount of computational time and the spatial pyramid originally introduced in [32] requires significant amount of memory (Approximately 250GB of memory is required to store the extracted features of 1 dataset out of 4 in the CUHK dataset).

3. CONCLUSION

In this paper, we compare generic, hand-crafted and aesthetics descriptors to examine the classification performance of image aesthetics. We have shown that basic generic descriptors can perform as well as the state of the art hand-crafted global descriptors. However, both generic and hand-crafted features are limited in terms of classification when we do not consider the spatial distribution of features. Spatial pyramid and subject region extraction are the main factors that lead to high classification accuracies in the literature. But they require significant amount of computational time and memory. In our future work, we will examine the correlation between the extracted features and overall image aesthetics. Instead of directly feeding the features to classifiers, we will focus on the individual relationships between the features and aesthetics to model a no-reference image aesthetics metric based on deep learning. We plan to use AVA dataset introduced in [21] to evaluate the aesthetics metrics because of the image variety, aesthetics scores and rich annotations.

4. REFERENCES

[1] R. Datta and J. Z. Wang, “Studying Aesthetics in Photographic Images,” 2006.

[2] D. Hasler and S. Sabine, “Measuring colourfulness in natural images,” 2003.

[3] G. Yildirim, A. Shaji, and S. Süsstrunk, “Estimating beauty ratings of videos using supervoxels,” *Proceedings of the 21st ACM international conference on Multimedia - MM ’13*, pp. 385–388, 2013.

[4] S. Dhar, T. L. Berg, and S. Brook, “High Level Describable Attributes for Predicting Aesthetics and Interestingness,” 2011.

[5] Y. Luo and X. Tang, “Photo and Video Quality Evaluation : Focusing on the subject,” pp. 386–399, 2008.

[6] A. Levin, “Blind Motion Deblurring Using Image Statistics,” 2006.

[7] X. Tang, W. Luo, and X. Wang, “Content-Based Photo Quality Assessment,” *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.

[8] A. K. Moorthy, P. Obrador, and N. Oliver, “Towards Computational Models of Visual Aesthetic Appeal of Consumer Videos,” 2010.

[9] S. Bhattacharya, R. Sukthankar, and M. Shah, “A framework for photo-quality assessment and enhancement based on visual aesthetics,” *Proceedings of the international conference on Multimedia - MM ’10*, p. 271, 2010.

[10] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, “Assessing the aesthetic quality of photographs using generic image descriptors,” pp. 1784–1791, 2011.

[11] A. Oliva and A. Torralba, “Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope ,” vol. 42, no. 3, pp. 145–175, 2001.

[12] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, and D. Maupertuis, “Visual Categorization with Bags of Keypoints,” *ECCV SLCV Workshop*, 2004.

[13] J. Sivic and A. Zisserman, “Video Google: a text retrieval approach to object matching in videos,” *Proceedings Ninth IEEE International Conference on Computer Vision*, , no. Iccv, pp. 1470–1477 vol.2, 2003.

[14] Photo.net, “http://photo.net,” .

[15] DPChallenge, “http://www.dpchallenge.com,” .

[16] R. Datta, J. Li, and J.Z. Wang, “Algorithmic inferencing of aesthetics and emotion in natural images: An exposition,” in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, 2008, pp. 105–108.

[17] Alipr, “http://alipr.com,” .

[18] Terragalleria, “http://www.terragalleria.com,” .

[19] H. Mller, P. Clough, Th. Deselaers, and B. Caputo, *ImageCLEF: Experimental evaluation in visual information retrieval series. The information retrieval series*, Springer, 2010.

[20] University of Florida, “International affective picture system,” .

[21] N. Murray, D. Barcelona, L. Marchesotti, and F. Perronnin, “AVA : A Large-Scale Database for Aesthetic Visual Analysis,” 2012.

[22] Y. Ke, X. Tang, and F. Jing, “The Design of High-Level Features for Photo Quality Assessment,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 1, pp. 419–426.

[23] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang, “Classification of digital photos taken by photographers or home users,” *Proceedings of Pacific Rim Conference on Multimedia*, 2004.

[24] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[25] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in *Proc. BMVC*, 2002, pp. 36.1–36.10, doi:10.5244/C.16.36.

[26] A. Vedaldi and B. Fulkerson, “VLFeat: An open and portable library of computer vision algorithms,” 2008.

[27] R. Khan, J. Van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat, “Discriminative color descriptors,” 2013.

[28] J. Van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, “Learning color names for real-world applications.,” *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 18, no. 7, pp. 1512–23, July 2009.

[29] R. Benavente, M. Vanrell, and R. Baldrich, “Parametric fuzzy sets for automatic color naming.,” *Journal of the Optical Society of America. A, Optics, image science, and vision*, vol. 25, no. 10, pp. 2582–93, Oct. 2008.

[30] J. Van De Weijer and C. Schmid, “Coloring local feature extraction,” in *In ECCV, 2006. MENSINK et al.: TMRF FOR IMAGE AUTOANNOTATION*.

[31] J. Van de Weijer, “http://cat.uab.es/ joost/software.html,” .

[32] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2006, CVPR ’06, pp. 2169–2178, IEEE Computer Society.