Assessing the aesthetic quality of photographs using generic image descriptors

Luca Marchesotti, Florent Perronnin, Diane Larlus, Gabriela Csurka
Textual and Visual Pattern Analysis (TVPA)
Xerox Research Centre Europe (XRCE), France
{luca.marchesotti,florent.perronnin,diane.larlus,gabriela.csurka}@xerox.com

Abstract

*In this paper, we automatically assess the aesthetic properties of images. In the past, this problem has been addressed by hand-crafting features which would correlate with best photographic practices (*e.g. *"Does this image respect the rule of thirds?") or with photographic techniques (*e.g. *"Is this image a macro?"). We depart from this line of research and propose to use generic image descriptors to assess aesthetic quality. We experimentally show that the descriptors we use, which aggregate statistics computed from low-level local features, implicitly encode the aesthetic properties explicitly used by state-of-the-art methods and outperform them by a significant margin.*

1. Introduction

The objective of image quality assessment is to design methods which can automatically predict the perceived quality of an image [29]. Aesthetic models have many applications of practical value. In image retrieval systems, similar images (from a content-based perspective) could be re-ranked using aesthetic properties. They could help a user to select the best pictures from his collection to make photo-albums. Also, these models could be deployed directly in photo cameras to make real-time suggestions.

Image quality assessment can be seen as a binary classification problem: is this image "good" or "bad"? It is an intrinsically challenging problem for several reasons. Firstly, visual data is very rich and ambiguous. Secondly, when judging photographs, people are often confronted to personal tastes. Finally, if one might agree that low level degradations (*e.g.* out of focus image) are - in general - an indicator of poor quality, it is more difficult to find a consensus on higher level visual properties such as color harmonies, layout, lighting conditions, etc. With all these difficulties, one might even question the possibility to learn generic models encoding photographic preference.

However, there is some agreement between professionals, about some best photographic practices (*e.g.* rule of thirds) as well as photographic techniques (*e.g.* macro). Based on this knowledge, previous work on image quality assessment has proposed hand-crafted visual descriptors, to mimic these photographic rules. Combined with supervised classifiers, they have achieved good results in predicting image quality. Despite the fact that the described hand-crafted features are aesthetically motivated, they have some disadvantages i) they are non exhaustive: they can never cover all possible photographic principles ii) they are computational expensive, at least for the most successful ones, and iii) they use heuristics, which may not generalize well to similar applications.

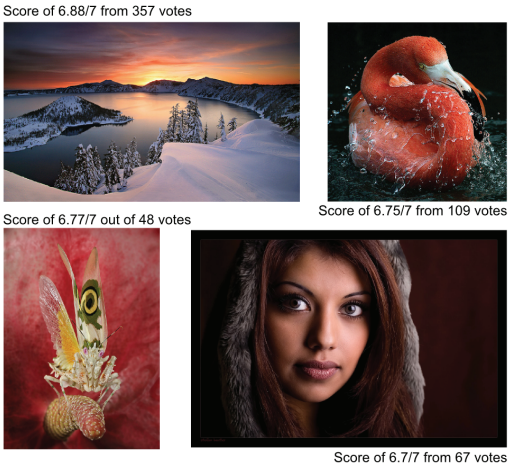

Figure 1. Photos highly rated by peer voting in an on-line photo sharing community (*photo.net*).

That is why we pursue a different direction and we consider some well-designed but more *generic image features*, which have been successfully used for semantic tasks. While handcrafted methods try to model best practices *explicitly*, we think that the same information is already, at least partially, *implicitly* modeled in generic image features. We show in this paper that these task-independent descriptors successfully capture aesthetic properties too, and outperform state-of-the-art features, on two different aesthetic

978-1-4577-1102-2/11/$26.00

datasets. Qualitative results confirm our initial intuition that generic descriptors are able to learn visual properties in an efficient and implicit manner.

In particular, we look at two families of generic descriptors, the Bag-of-Visual-Words (BOV [2]) and the Fisher Vector (FV [11]), which encode the distribution of local statistics. Such descriptors have been proven to perform well for a large set of semantic tasks. Since aesthetic classification is a high-level vision task, in the same manner that object classification/detection is a high-level semantic task, it should not be surprising that the same features work well for both problems. Also, a convenient aspect of these descriptors is their nice scaling properties to large datasets.

The remainder of this article is organized as follows. In the next section, we review related work. In section 3, we describe the generic content-based features we use in this paper. Section 4 describes the datasets considered in the experiments. In section 5, we compare previous aesthetic features with generic content-based features on these datasets. We finally propose a discussion.

2. Related Work

Most recent quality assessment techniques are based on a learning approach, where a prediction function is learned from a labeled training set [7, 26, 19]. The focus is currently on the design of the appropriate features for this learning problem. These features should capture the visual properties that make a specific image appealing for the majority of the viewers. Typically, *low* and *high level* features are jointly used to perform this kind of analysis. Low-level features include the exposure, contrast, colorfulness and texture of the photographs. High-level features are related to the analysis of image's layout.

Early works employed low level features inspired by perceptually motivated metrics [4, 12, 3]. The most popular ones [24] involve simple statistics (*e.g.* mean and standard deviation) evaluated over the entire image to characterize signal-level degradations such as random or structured noise (*e.g.* salt and pepper noise, jpeg artifacts, ringing) [23]. Other descriptors are based on blur estimation techniques where the blur is modeled with a Gaussian smoothing filter [27].

Currently, researchers are investigating features at a higher level of abstraction. This follows the assumption that high quality images adhere to several photographic rules and best practices of professional photographers (*e.g.* out of focus background, chiaroscuro lighting, motion blur, high speed photography, rule of thirds, macro/close-up, leading lines, altered viewpoint) [13].

This is the reason why, in [26] handcrafted features are designed to detect such photographic rules. Segmentation techniques are often used to attain an higher level description of the image, *e.g.* for the rule of thirds, you need to determine the location of the foreground object. For the segmentation, a simple k-means region segmentation could be used [4], while more complex methods can spot in focus objects by detecting blurry edges (corresponding to background) [17]. This latter method is particularly useful to detect images shot with a popular photographic technique involving high aperture settings. Another approach is to locate specific objects, *e.g.* [15] uses classic techniques (Viola-Jones face detection) to detect face ovals and locate eyes and mouth.

The shape of the segmented objects can be characterized through geometric contexts and region analysis algorithms [26]. Also, the absolute position of these objects could be evaluated against locations within the image that are considered as important from an aesthetic point of view (*e.g.* points individuated by the rule of thirds, golden ratio [13]). In addition, metrics are defined [19] to describe pleasant layouts in terms of relative position of objects, presence of vanishing points or perspectives.

In this paper, we deviate from this line of research, and we propose to use *generic image descriptors* [2, 20]. The proposed image descriptors have been successfully applied to *semantic tasks* such as object/scene retrieval [25, 21], image classification/annotation [2, 22] and object localization [10]. They have proven to be versatile, working in scenarios with large number of classes, and scaling to large datasets. In the next section, we describe two popular kinds of generic image descriptors.

3. Generic Image Features

We propose to use of a generic content-based local image signature for aesthetic assessment. We first consider the the Bag-Of-Visual-words descriptor (BOV, also called bag-of-features) [2, 25] which is probably the most widely used for semantic tasks. We then use one of its recent extension, the Fisher Vector (FV, [20, 22]), that was shown to yield state-of-the-art results for tasks such as image retrieval and image classification. For both descriptors we encode gradient information using SIFT [16] and color information.

Our motivation to use these descriptors is the following: rather than trying to encode photographic rules explicitly, we encode them implicitly in generic content-based features such as the BOV or the FV which describe the distribution of local patches within the image (BOV: discrete distribution, FV: continuous distribution). Indeed, each patch (*e.g.* SIFT or color) can tell us a lot about the local properties of an image (*e.g.* "Does this patch contain sharp edges?" or "Is the color of this patch saturated?"). Moreover, by aggregating patch-level information into an image-level BOV or FV, one can take into account the global composition (*e.g.* "Do we have a mix of sharp patches and blurry ones?" or "Is there a dominant color or a mixture of colors in this image?"). Finally, using the spatial pyramid framework [14],

we can model the layout (*e.g*. rule of thirds).

We also consider the GIST descriptor [18], originally used for scene categorization, since it should also capture the layout of images. Below we give a quick description of these descriptors, more details can also be found in [18, 2, 20].

GIST. Oliva and Torralba [18] introduced the GIST descriptor as a low-dimensional scene descriptor. A set of perceptual dimensions, that represent the global structure of a scene (naturalness, openness, roughness, expansion, ruggedness) is estimated using spectral information and coarse localization. In practice, an image is partitioned in a 4×4 regular grid and a 20-D histogram of gradients is computed for each region, and for each color channel. The concatenation of all histograms produces a 960-D vector.

Bag-Of-Visual Words (BOV). In the BOV representation [25, 2] an image is described by a histogram of quantized local features. More precisely, an unordered set of local patches are first extracted and described, for instance by SIFT descriptors [16]. A visual vocabulary, *i.e*. a set of prototypical features, is learned by clustering a large number of local descriptors. The set of local features extracted from a given image is then transformed into a fixed-length histogram representation by counting the number of local descriptors assigned to each visual word. The BOV was shown to be successful in applications such as image retrieval [25] and classification [2].

In this work, we follow [9] and use a Gaussian Mixture Model (GMM) to model the distribution of local features, *i.e*. we have a probabilistic visual vocabulary. The GMM vocabulary provides a principled way to cope with assignment uncertainty as each local feature is assigned with a probability to all visual words. We finally square-root the histograms as suggested in [22, 28], which corresponds to an explicit embedding of the data in the case of the Bhattacharyya kernel. We verified experimentally that it did improve the results. While, in its original formulation, the BOV does not contain any layout information, Lazebnik *et al*. [14] proposed to include coarse spatial information by dividing hierarchically the image into a set of regions, computing one BOV histogram per region and then concatenating these region-level representations.

Fisher Vector (FV). The FV [11, 20] extends the BOV by going beyond counting (0-order statistics) and by encoding statistics (up to the second order) about the distribution of local descriptors assigned to each visual word. It was shown to outperform the BOV descriptor in applications such as image classification [20, 22] and image retrieval [21]. This can be easily understood as it treats images as continuous distribution while BOV treats them as discrete distributions. A significant advantage with respect to the BOV is that high-dimensional discriminative signatures can be obtained even with small vocabularies, and therefore at a low CPU cost.

The FV $\mathcal{G}_\lambda^X$ characterizes a sample $X = \{x_t, t = 1 \dots T\}$ by its deviation from a distribution u_λ (with parameters λ):

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X. \tag{1}$$

G_λ^X is the gradient of the log-likelihood with respect to λ:

$$G_\lambda^X = \frac{1}{T} \nabla_\lambda \log u_\lambda(X). \tag{2}$$

L_λ is the Cholesky decomposition of the inverse of the Fisher information matrix F_λ of u_λ, *i.e*. $F_\lambda^{-1} = L'_\lambda L_\lambda$ where by definition:

$$F_\lambda = E_{x \sim u_\lambda} \left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)' \right]. \tag{3}$$

In our case, X is the set of T local descriptors extracted from an image and $u_\lambda = \sum_{i=1}^N w_i u_i$ is a GMM which models the generative process of local descriptors (*i.e*. the probabilistic visual vocabulary).

As shown in [22], square-rooting and L2-normalizing the FV can greatly enhance the classification accuracy. Also, rough spatial layout information can be incorporated in a manner similar to the BOV: one can split an image into several regions, compute one FV per region and concatenate the per-region FVs.

4. Datasets

The image datasets which are used to study aesthetics typically consist of photographic images shared on social networks. Communities like *photo.net*[1], *DPChallenge*[2] or *Terra Galleria*[3] gather a large number of expert and amateur photographers who share, view and judge photos online. These photographers also agree on the most appropriate annotation policy to score the images. Such policies can include textual labels ("like it", "don't like it") or a scale of numerical values (ratings). From these annotations, images can be labeled as being visually appealing or not. Such a ground truth allows a fair quantitative evaluation of the different methods.

For our evaluation, we used the two public databases described below.

4.1. Photo.net dataset

Photo.net is a community network where registered users rate images with a score between 1 (Ugly) and 7 (Beautiful). The users are provided by the site administrators with the following guidelines for judging images: *"Reasons for a rating closer to 7: a)it looks good, b)it attracts/holds attention, c)it has an interesting composition,*

[1] http://www.photo.net
[2] http://www.dpchallenge.com
[3] http://www.terragalleria.com

d)it has great use of color, e)(if photojournalism) contains drama, humor, impact, f)(if sports) peak moment, struggle of athlete". Figure 1 shows sample photos of high quality with their scores and number of votes.

For our evaluation, we used two different sets of images derived from *photo.net* that we call Photo.net (PN) and Photo.Net cropped (PNc).

PN was introduced in [4] and used in [4, 6]. It consists of 3,581 images. However, only URLs of the original images are provided. Since many images were removed from the website by their owners, we are left with 3,118 images. At visual inspection of PN, we have noticed a correlation between images receiving a high grade and the presence of frames manually created by the owners to enhance the visual appearance (see examples in Figure 2). In particular, we manually detected that more than 30% of the images are framed. Since this introduces a bias in the database, we manually removed all the borders and we created a second dataset that we called PNc (PN cropped).

Figure 2. Sample images from PN with borders manually created by photographers to enhance the photo visual appearance.

4.2. CUHK dataset

The second set of images we consider in our experiments is derived from the web site *DPChallenge.com*. We use the CUHK collection created by the Chinese University of Hong Kong [12] from DPChallenge where 60,000 photos were extracted, choosing the ones rated by at least a hundred users. The photos average rating is used as the ground truth, and the highest and lowest 10% average rates have been assigned to the "good" and "bad" classes respectively and manually assessed. For this reason, this dataset is less challenging than photo.net.

12,000 images are available. Ke *et al*. [12] observed the same bias for images with border as we did for PN, so they removed all the frames from the images they released.

5. Experimental Validation

We now validate our initial hypothesis: generic content-based descriptors are useful for aesthetic prediction. We first describe the experimental set-up. We then report detailed experiments on the smaller PN dataset. We finally report results on CUHK.

5.1. Experimental Set-Up

State-of-the-art features. We have implemented the 56 aesthetic features of Datta *et al*. [4], the 7 features of Ke *et al*. [12] and the 5 features of Luo *et al*. [17]. We managed to replicate the results reported in [4] and [12] using the same experimental protocols and datasets. However, we did not manage to replicate the results reported by [17]. Indeed, while [17] claims a reduction of the classification error of 80% with respect to [4], our re-implementation of [17] actually leads to results which are significantly worse than those of [4]. Therefore, we do not include the features of [17] in our baseline aesthetic features.

Generic Content-Based Features. To compute GIST features we use the code made available online by the authors of [18]. For BOV and FV, local patches of size 32×32 are extracted regularly on grids every 4 pixels at 5 scales. We use two types of local descriptors to represent the patches: SIFT and Color descriptors. SIFT divides the local patch in a 4x4 grid, and it computes an histogram of oriented gradients in each bin of the grid. Similarly the color descriptor we use divides the patch into a 4x4 grid and it computes simple statistics per color channel for each bin of the grid. This produces 128-dimensional SIFT descriptors [16] and 96-dimensional color descriptors [22]. Both are reduced with PCA to 64 dimensions. The probabilistic visual vocabulary, *i.e*. the GMM, is learned using a standard EM algorithm. For the BOV we use a GMM with 1,024 Gaussians and for the FV a GMM with 256 Gaussians. For the spatial pyramid, we follow the splitting strategy adopted by the winning system of PASCAL VOC 2008 [8]. We extract 8 vectors per image: one for the whole image, three for the top, middle and bottom regions and four for each of the four quadrants. The hope is that the pyramid can encode information about the image composition.

Classification. We learn linear SVMs with a hinge loss using the primal formulation and a Stochastic Gradient Descent (SGD) algorithm [1]. For the BOV and FV, we run two separate systems, one for SIFT and one for color features. The combined version is simply obtained with late fusion, *i.e*. by averaging the scores of the two systems.

5.2. Evaluation protocol

PN. In agreement with [4], which first used the photo.net dataset, we compute for each image i, the average of all scores available $q_{av}(i)$. A statistical analysis of the scores can be found in [5] and it shows that in *photo.net* ratings are skewed towards positive values and consensus among users is generally low. Also, they observed the value 5.0 to be median. Following [5], we set two thresholds $\theta_1 = 5 + \delta/2$ and $\theta_2 = 5 - \delta/2$. We then annotate each image with the label "good" if $q_{av}(i) \geq \theta_1$ and "bad" if $q_{av}(i) \leq \theta_2$. δ is used to artificially create a gap between high and low quality images, as pictures lying in this gap are likely to represent noisy data in the peer-rating process. As in [4] we vary this δ value in our experiments. Increasing the value δ obviously

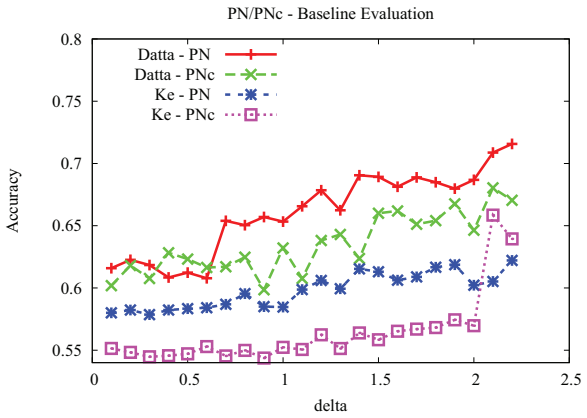

Figure 3. Classification accuracy for the Datta *et al*. [4] and the Ke *et al*. [12] features, evaluated on PN and PNc after removing the photo frames (PNc).

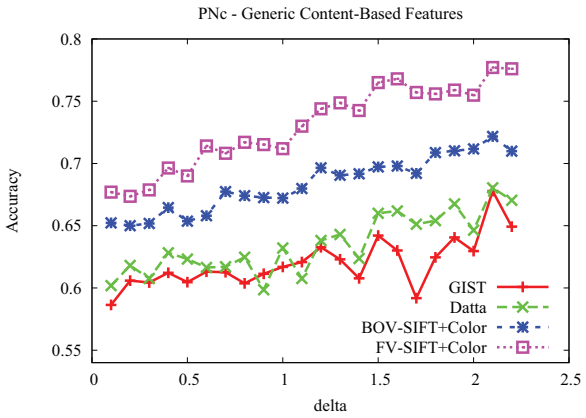

Figure 4. Classification accuracy on PNc using GIST, and combined (SIFT+Color) version of BOV and FV descriptors.

makes the task easier. Please note that images belonging to the "bad" class are not really bad per se. They only correspond to images that received lower rates. We perform 5-fold cross-validation as suggested in [4] and report the average results.

CUHK. Following [12], we used half of the images (6,000) for training, and half (6,000) for testing. In this case, we do not have access to ratings therefore we rely on the binary labels ("good" and "bad") provided by the authors.

5.3. Quantitative Evaluation

Baseline Features. In Figure 3 we plot the classification accuracy as a function of the rating threshold δ for Datta *et al*. [4] and Ke *et al*. [12] features. To assess the bias introduced by the ornament frames we performed the evaluation on the 3118 original images of the PN dataset, and on the same images cropped manually. On average, the bias accounts for a couple of accuracy points in both features set. We observe that Datta features outperform Ke features. We also tried to merge the Datta and Ke features but no significant improvement was achieved. We notice that all previous results published on [4] and [5] do not take into account the advantage provided by the frames of high quality images. Therefore, in the rest of the experiments we take as a reference datasets PN with the cropped frames (PNc) and as baseline the features proposed by [4] (referred to as "Datta" features).

Proposed Generic Content-Based Features. We report in Figure 4 results with the 3 proposed content-based features: GIST, BOV and FV (for both BOV and FV we extract SIFT and color based descriptors and we combined the results in late fusion). One can draw the following conclusions.

First, even the GIST descriptor performs somewhat comparably to the Datta features which shows that simple general purpose content-based features which were not designed for aesthetic assessment can perform as well as features which were hand-crafted for this specific problem.

Second, the FV and the BOV outperform GIST and Datta showing that the distribution of local descriptors has a high discrimination potential for image quality assessment. In the image classification literature, the visual vocabulary is often said to be an intermediate representation which bridges the *semantic gap* between the low-level local features and high-level semantic concepts. It seems that it also helps bridging the *aesthetic gap* between the low-level features and the high-level perception.

Third, overall the FV is the best performing descriptor. It was already observed for several semantic tasks, like object categorization [20] and image retrieval [21], that the FV outperforms BOV, because it models the distribution of local descriptors in a continuous manner.

We now focus on the FV representation, the results are shown in Figure 5. The two classifiers trained with color and SIFT features have equivalent performances. We also observe that the spatial pyramid has very little impact on the FV. This can have two reasons. We may have an inadequate partitioning for the task (different region choices could be tried). Also, the pyramid increases significantly the dimensionality of the image features, so the dataset may be too small to train reliably a classifier with these descriptors.

CUHK results. We complete this quantitative evaluation by experimenting on the second reference dataset CUHK (see Table 1). First, the GIST descriptor performs significantly worse than all other descriptors on this dataset. Second, both state-of-the-art feature sets perform equally. Again the best performing feature is the FV. For both the BOV and the FV, the color feature works better, but difference is stronger for FV. We could not get any improvement with the late

Features	GIST	Datta	Ke	BOV-SIFT-SP	BOV-Color-SP	FV-SIFT-SP	FV-Color-SP
Accuracy	67.96	75.85	76.53	81.36	81.86	82.13	89.90

Table 1. Performances on the CUHK dataset for Datta's and Ke's features, BOF and FV descriptors.

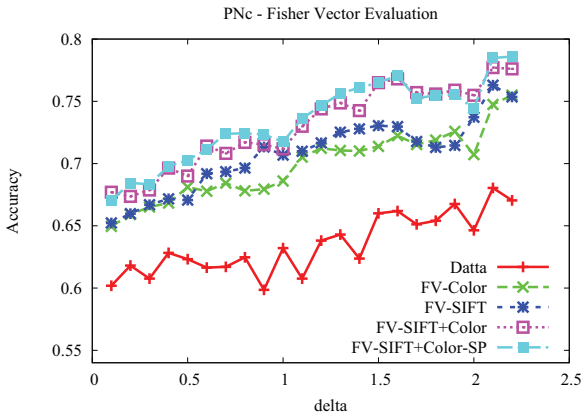

Figure 5. Combination of the different flavors of FV on PNc. SIFT and Color are considered individually, combined, and used in the spatial pyramid (SP) framework.

fusion of SIFT and color features on this dataset.

6. Discussion

The good results obtained by our method, which uses generic image descriptors, show evidences that the proposed strategy is able to capture a wealth of statistics useful for aesthetic evaluation of photographs. In this section, we propose some qualitative results, and we try to give hints on the reason of this success. Drawing conclusions from the visual inspection of our results is a difficult task, as the datasets are large, and the labels extracted from the scores are available without any explanation. However, we would like to discuss a few trends that emerged during this qualitative evaluation. To comment the visual results, we look at the photographic practices and techniques known to produce aesthetically pleasing results.

We focus on the CUHK dataset since it contains more images, and more user votes in comparison with PN. However our observations still hold for PN. We take into account two of the most performing features: the FV-Color-SP and FV-SIFT-SP descriptors. Figure 7 shows the 24 images receiving the highest scores, and the lowest scores, for both descriptors. The ground truth labels are displayed in the form of colored frames (green for images rated as good, and red for the others).

For the SIFT based descriptor, the most obvious observation is that the blur information is well encoded. All images with high classification scores are very sharp with strong contrasted edges. Similarly, all blurry images get very low scores, as well as low resolution images, where edges cannot be successfully extracted (see the first false negatives *i.e.* positive images with the lowest score, in Figure 6). The scores remain very high for photos with sharp foreground and out-of-focus background. More generally, the preference of the SIFT-based classifier goes to images with a clearly identified foreground with dark, uniform or out-of-focus background. Conversely, cluttered images are classified as negative. Finally, high dynamic range photos are also ranked high. We think all these aesthetic properties are captured by the SIFT descriptor which encodes local shape, texture and edges, but also illumination. If we look at failures for SIFT-based descriptors in Figure 6, the first false positive images correspond to images whose low-level quality is good, but whose subject may not be considered as interesting.

Now let us look at the color-based generic descriptors. The chromatic properties of images are, as expected, what differentiate high quality and low quality images the most. A lot of sunset images, in general quite popular, are within the top images. Typically, such images have one dominant color, or complementary colors (*e.g.* red and green, blue and yellow). In contrast, images with too many colors, or cheerless colors, received low scores. But color descriptors also learnt successfully what makes a black and white or a sepia tone image visually appealing, and a number of them obtain a high score.

Even if the observation is not as strong as for SIFT, color descriptors also capture the blur information. Sharp and well contrasted images, get higher scores, while blurry images do not. Color descriptors are based on color visual words, but they include some local geometry (4x4 grid) too. This means that they can extract (even though not as well as SIFT) some information linking to edges, contrast and sharpness. This duality would explain why color descriptors outperform SIFT.

As shown in Figure 5, the spatial pyramid does not seem to bring a significant improvement. We note however that for both low level descriptors, the corresponding FV gives a high score to images with an artistic layout. Images following the rules of third, exhibiting an off-center subject, or strong leading lines seem to get a higher classification score than others. This could be because the layout properties correlate with other aesthetic aspects.

All these observations are in favor of our original claim: *generic descriptors are able to learn in an efficient and implicit manner what state-of-the-art descriptors have been*

Figure 6. (top) negative images with higher scores, (bottom) positive images with lowest scores, for FV-Color-SP and FV-SIFT-SP.

trying to encode explicitly. For this reason, we strongly believe that they are well-suited for the quality assessment of photographs.

For our experiments we considered the two largest publicly available aesthetic datasets. We believe that, because the descriptors we consider are high dimensional, they would strongly benefit from large scale datasets. In future work, we will address the challenges raised while building a large quality assessment dataset.

References

[1] L. Bottou. Sgd. http://leon.bottou.org/ projects/sgd. 4

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV SLCV Workshop*, 2004. 2, 3

[3] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik. Image quality assessment based on a degradation model. *Image Processing, IEEE Transactions on*, 9(11):363–650, Apr. 2000. 2

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006. 2, 4, 5

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, 2008. 4, 5

[6] R. Datta, J. Li, and J. Z. Wang. Learning the consensus on visual quality for next-generation image management. In *ACM MM*, 2007. 4

[7] R. Datta and J. Z. Wang. Acquine: aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *MIR*, 2010. 2

[8] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, 2008. 4

[9] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005. 3

[10] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, sep 2009. 2

[11] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999. 2, 3

[12] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006. 2, 4, 5

[13] Kodak. *How to take good pictures : a photo guide*. Random House Inc, 1982. 2

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 3

[15] C. Li, A. Gallagher, A. Loui, and T. Chen. Aesthetic visual quality assessment of consumer photos with faces. In *ICIP*, 2010. 2

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2, 3, 4

[17] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008. 2, 4

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001. 3, 4

[19] N. O. Pere Obrador, L. Schmidt-Hackenberg. The role of image composition in image aesthetics. In *ICIP*. 2010. 2

[20] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 2, 3, 5

[21] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010. 2, 3, 5

[22] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 2, 3, 4

[23] H. Sheikh, A. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *Image Processing, IEEE Transactions on*, 14(11):1918–1927, 2005. 2

[24] H. Sheikh, M. Sabir, and A. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Transactions on*, 15(11):3440–3451, 2006. 2

[25] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 2, 3

[26] M. S. Subhabrata Bhattacharya, Rahul Sukthankar. A coherent framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM MM*, 2010. 2

[27] H. Tong. Blur detection for digital images using wavelet transform. In *In Proceedings of IEEE International Conference on Multimedia and Expo*, pages 17–20, 2004. 2

[28] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010. 3

[29] Z. Wang, H. R. Sheikh, and A. C. Bovik. *The Handbook of video databases: design and applications*, chapter 41. CRC press, 2003. 1

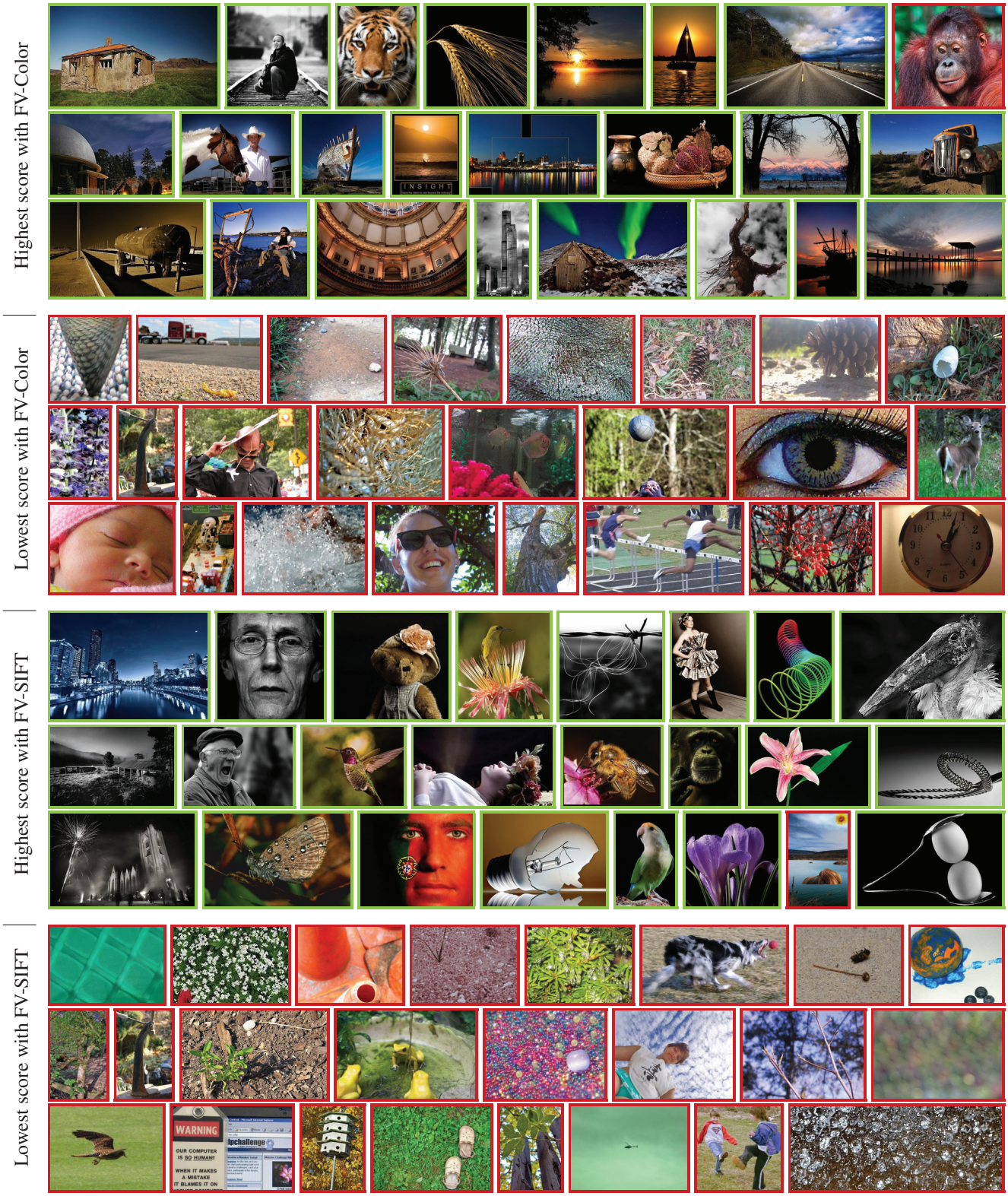

Figure 7. Qualitative results on the CUHK dataset: For FV-Color-SP and FV-SIFT-SP, the highest and lowest rank images are shown. The colored frame represents the ground truth (green for "good" and red for "bad"). The figure is best viewed in color