High Level Describable Attributes for Predicting Aesthetics and Interestingness

Sagnik Dhar Vicente Ordonez Tamara L Berg
Stony Brook University
Stony Brook, NY 11794, USA
tlberg@cs.stonybrook.edu

Abstract

With the rise in popularity of digital cameras, the amount of visual data available on the web is growing exponentially. Some of these pictures are extremely beautiful and aesthetically pleasing, but the vast majority are uninteresting or of low quality. This paper demonstrates a simple, yet powerful method to automatically select high aesthetic quality images from large image collections.

Our aesthetic quality estimation method explicitly predicts some of the possible image cues that a human might use to evaluate an image and then uses them in a discriminative approach. These cues or high level describable image attributes *fall into three broad types: 1) compositional attributes related to image layout or configuration, 2) content attributes related to the objects or scene types depicted, and 3) sky-illumination attributes related to the natural lighting conditions. We demonstrate that an aesthetics classifier trained on these describable attributes can provide a significant improvement over baseline methods for predicting human quality judgments. We also demonstrate our method for predicting the "interestingness" of Flickr photos, and introduce a novel problem of estimating query specific "interestingness".*

1. Introduction

Automating general image understanding is a very difficult and far from solved problem. There are many subproblems and possible intermediate goals on the way toward a complete solution, including producing descriptions of what objects are present in an image (including their spatial arrangements and interactions), what general scene type is shown (*e.g.* a beach, office, street etc.), or general visual qualities of the image (such as whether a picture was captured indoors, or outside on a sunny day). While none of these are solved problems either, progress has been made in the research community toward partial solutions.

In this paper we build on such progress to develop techniques for estimating high level describable attributes of images that are useful for predicting perceived aesthetic quality of images. In particular we demonstrate predictors for:

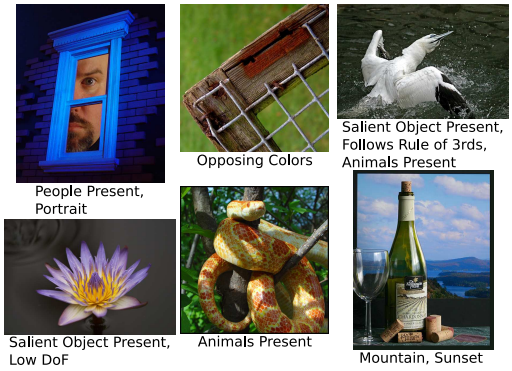

Figure 1. High Level describable attributes automatically predicted by our system.

1. Compositional Attributes - characteristics related to the layout of an image that indicate how closely the image follows photographic rules of composition.
2. Content Attributes - characteristics related to the presence of specific objects or categories of objects including faces, animals, and scene types.
3. Sky-Illumination Attributes - characteristics of the natural illumination present in a photograph.

We use the phrase *high level describable attributes* to indicate that these are the kinds of characteristics that a human might use to describe an image. Describability is key here so that we can ask people to label images according to the presence or absence of an attribute and then use this labeled data to train classifiers for recognizing image attributes.

Recent work on attributes for faces has shown that for face verification, describable facial attributes can produce better performance than purely low level features [12]. While our focus is on attributes of images not of faces, we pursue a similar direction to demonstrate the power of attributes for: estimation of aesthetic quality (Sec 3.1), estimation of general interestingness (Sec 3.2), and a new problem of estimation query specific interestingness (Sec 3.3).

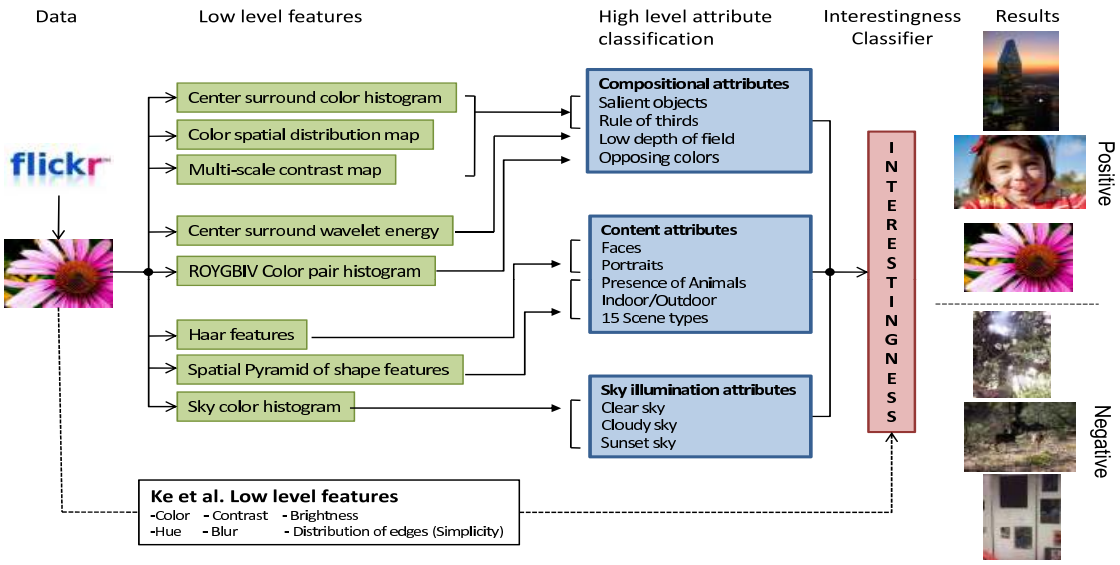

Figure 2. Overview of our method for estimating interestingness (aesthetic quality follows a similar path). From left to right: a) example input image b) low level features are estimated c) high level attributes are automatically predicted by describable attribute classifiers, d) interestingness is predicted given high level attribute predictions (or optionally in combination with level features [11] – dashed line).

While much previous work on aesthetics prediction has provided intuition about what high level attributes might be useful, they have used this intuition to guide the design of relevant low level image features. Our approach, on the other hand explicitly trains classifiers to estimate describable attributes and evaluates the accuracy of these estimates. Furthermore, we demonstrate that classifiers trained on high level attribute predictions are *much more effective* than those trained on purely low level features for aesthetics tasks, and can be made even more accurate when trained on a combination of low level features and high level attributes (fig 5). Our other main contributions include a focus on extracting high level visual attributes of images (as opposed to objects), and novel attributes related to image layout.

1.1. Previous Work

Our work is related to three main areas of research: estimating visual attributes, estimating the aesthetics of photographs, and human judgments of aesthetics.

Attributes: Recent work on face recognition has shown that the output of classifiers trained to recognize attributes of faces – gender, race, age, etc. – can improve face verification [12]. Other work has shown that learning to recognize attributes can allow recognition of unseen categories of objects from their description in terms of attributes, even with *no* training images of the new categories [13, 5, 7]. Our work is related to these methods, but while they focus on attributes of objects (*e.g.* "blond" person, or "red" car), we look at the problem of extracting high level describable attributes of images (*e.g.* "follows rule of 3rds").

Aesthetics: There has been some previous work on estimating the aesthetic quality of images, including methods to differentiate between images captured by professional photographers versus amateurs [24, 11, 25, 3, 23, 17]. This prior work has utilized some nice intuition about how people judge aesthetic quality of photographs to design low level features that might be related to human measures. Datta et al select visual features based on artistic intuition to predict aesthetic [3] and emotional quality [4]. Tong et al use measures related to the distortion [25]. Ke et al select low level features such as average hue, or distribution of edges within an image, that may be related to high level attributes like color preferences or simplicity [11]. Our method is most similar to Luo & Tang [17], who also consider ideas of estimating the subject of photographs to predict aesthetic quality. The main difference between these approaches and ours is that instead of using human intuition to design low level features, we explicitly train and evaluate prediction of high level describable attributes. We then show that aesthetics classifiers trained on these attributes provides a significant increase in performance over a baseline method from Ke et al (fig 5). Our overall performance is comparable to the results reported in Luo & Tang [17], but seems to perform somewhat better, especially in the high precision low recall range – arguably the more important scenario for users trying to select high aesthetic quality photographs from large collections.

Human Judgment of Aesthetics: The existence of preferred views of objects has long been studied by Psychologists [20]. Photographers have also proposed a set of com-

position rules for capturing photos of good aesthetic quality. In some interesting recent work there have been several studies that expand the idea of view preferences to more general notions of human aesthetics judgment, including ideas related to compositional rules. These experiments include evaluating the role of color preferences [22, 18] and spatial composition [8, 1]. Other work in computational neuroscience has looked at developing models of visual attention including ideas related to saliency *e.g.* [10]. Some of our attributes are directly related to these ideas, including predicting the presence of opposing colors in images, and attributes related to the presence of salient objects, and arrangement of those objects at preferred locations.

1.2. Overview of Approach

The first phase of our work consists of producing high level image attribute predictors (Secs 2.1, 2.2, 2.3). We do this by collecting positive and negative example images for each attribute, picking an appropriate set of low level features, and training classifiers to predict the attribute. In each case labelers are presented with an image and asked to label the image according to some attribute, for example "Does this image follow the rule of thirds?". Possible answers are yes, not sure, or no and only images that are consistently labeled as yes or no are used for training.

Next we demonstrate that these high level attribute predictors are useful for estimating aesthetic quality (DPChallenge) and "interestingness" (Flickr). For each application, a set of training images is collected consisting of highly ranked images as positive examples and low ranked images as negative examples. A classifier is then trained using the output of the high level attribute predictors we developed as features and evaluated on held out data. We also show results for training classifiers using only low level features, and using a combination of low level features and our high level attributes. Results on aesthetics for DPChallenge are in Sec. 3.1 and interestingness for Flickr are in Sec. 3.2. Finally we show results on ranking for specific query interestingness in Sec. 3.3.

2. Describable attributes

We have developed high level describable attributes (examples in fig 1) to measure three types of image information: image composition (Sec 2.1), image content (Sec 2.2), and sky-illumination (Sec 2.3).

2.1. Compositional Attributes

Our compositional attributes address questions related to the arrangement of objects and colors in a photograph, and correspond to several well known photographic rules of composition. These compositional attributes are:

- Presence of a salient object – a photo depicting a large salient object, well separated from the background.

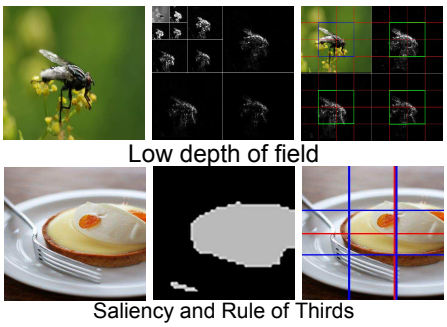

Figure 3. Example describable attribute computations. **Top** "Low DoF" (left: original image, center: wavelet transform, right: wavelet coefficients and center surround computation). **Bottom** "salient object presence" and "rule of 3rds" (left: original pic, center: detected salient object region, right: centroid and conformity to rule of 3rds).

- Rule of Thirds – a photo where the main subject is located near one of the dividing third-lines.
- Low Depth of Field – the region of interest is in sharp focus and the background is blurred.
- Opposing Colors – a photo that displays color pairs of opposing hues.

Presence of a salient object: We predict whether an image contains some large object, well separated from its background. To do this we take advantage of recent developments in automatic top down methods for predicting locations of salient objects in images [16]. As input image descriptors we implement 3 features related to saliency: a multi-scale contrast map, a center surround histogram map, and a center weighted color spatial distribution map – efficiently computing the features using integral image techniques [21]. All three of these feature maps are supplied to a conditional random field (CRF) to predict the location of salient objects (fig 3 shows a predicted saliency map). The CRF is trained on a set of images that contain highly salient objects. If an image does not contain a salient object, then the CRF output (negative of the log probability) will be high – estimated by the free energy value of the CRF.

We evaluate classification accuracy for this attribute using a set of 1000 images that have been manually labeled as to whether they contain a salient object. Precision-recall curves for predicting the presence of a salient object are shown in figure 4 (left plot, red), showing that our salient object predictor is quite accurate at estimating the presence of a salient object.

Rule of thirds: If you consider two vertical and two horizontal lines dividing the image into 6 equal parts (blue lines fig 3), then the compositional rule of thirds suggests that it will be more aesthetically pleasing to place the main subject of the picture on one of these lines or on one of their intersections. For our rule of thirds attribute, we again make use of the salient object detector. We calculate the

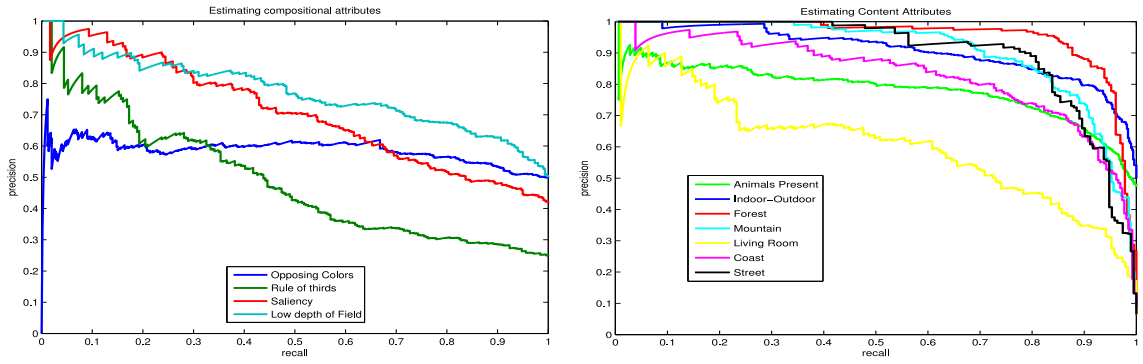

Figure 4. **Left:** Precision-Recall curves for some Compositional attributes: "Salient Object Present", "Follows Rule of 3rds", "Displays Opposing Colors". **Right:** Precision-Recall curves for some Content attributes: "presence of animals", "indoor-outdoor", various "scenes".

minimum distance between the center of mass of the predicted saliency mask and the 4 intersections of third-lines. We also calculate the minimum distance to any of the third-lines. We use the product of these two numbers (scaled to the range [0,1]) to predict whether an image follows the rule of thirds and evaluate this attribute on manually labeled images. Precision-recall curves are shown in fig 4 - left green.

Low depth of field: An image displaying a low depth of field (DoF) is one where objects within a small range of depths in the world are captured in sharp focus, while objects at other depths are blurred (often used to emphasize an object of interest). For our low DoF attribute we train an SVM classifier on Daubechies wavelet based features, indicative of the blurring amount [3]. The wavelet transform is applied to the image and then we consider the third level coefficients of the transformation in all directions (fig 3). Using a 4x4 grid over the image, we divide the sum of the coefficients in the four center regions by the sum of coefficients over all regions, producing a vector of 3 numbers, one for each direction of the transformation. A manually labeled dataset of 2000 images from Flickr and Photo.net is used to train and test our low DoF classifier. Precision-recall curves for the low DoF attribute are shown in fig 4 - left plot, cyan - demonstrating reliable classification.

Opposing colors: Some color singles, pairs, or triples are more pleasing to the eye than others [22, 18]. This intuition gives rise to the opposing colors rule which says that images displaying contrasting colors (those from opposite sides of the color spectrum) will be aesthetically pleasing. For this attribute we train classifiers to predict opposing colors using an image representation based on the presence of color pairs. We first discretize pixel values into 7 values. We then build a 7x7 histogram based on the percentage of each color pair present in an image and train an SVM classifier on 1000 manually labeled images from Flickr. Classification accuracy is shown in fig 4 - left plot, blue. Our classifier for this attribute is not extremely strong because even images containing opposing colors may contain enough other color noise to drown out the opposing color signal. However, this attribute still provides a useful signal for our aesthetics and interestingness classifiers.

2.2. Content Attributes

Content is often a large contributor to human aesthetic judgment. While estimating complete and accurate content is beyond current recognition systems, we present a set of high level content attributes that utilize current state of the art recognition technologies to predict:

- Presence of people – a photo where faces are present.
- Portrait depiction – a photo where the main subject is a single large face.
- Presence of animals – whether the photo has animals.
- Indoor-Outdoor classification – whether the photo was captured in an indoor setting,
- Scene type – 15 attributes corresponding to depiction of various general scene types (*e.g.* city, or mountain).

Presence of people & Portrait depiction: We use the Viola-Jones face detector [26] to estimate the presence of faces in an image (a proxy for presence of people). For this attribute we output a binary classification (1, if faces have been detected, and 0 otherwise). We manually label a test dataset of 2000 images from Photo.net and obtain 78.9% accuracy for predicting face presence. For portrait depiction we classify images as positive if they produce a face detection of size greater than 0.25 image size. We evaluate this feature on 5000 images from Photo.net hand labeled as portrait or non-portrait images and obtain 93.4% accuracy.

Object and scene attributes: For the "presence of animals", "indoor-outdoor classification" and "scene type" attributes we train 17 SVM classifiers using the intersection kernel computed on spatial pyramid histograms [14] (1 each for animals, and indoor-outdoor, and 15 for various scene categories). In particular we compute the SPM histograms on visual dictionaries of SIFT features [15] captured on a uniform grid with region size 16x16 and spacing of 8 pixels. The SIFT features for 100 random images are clustered

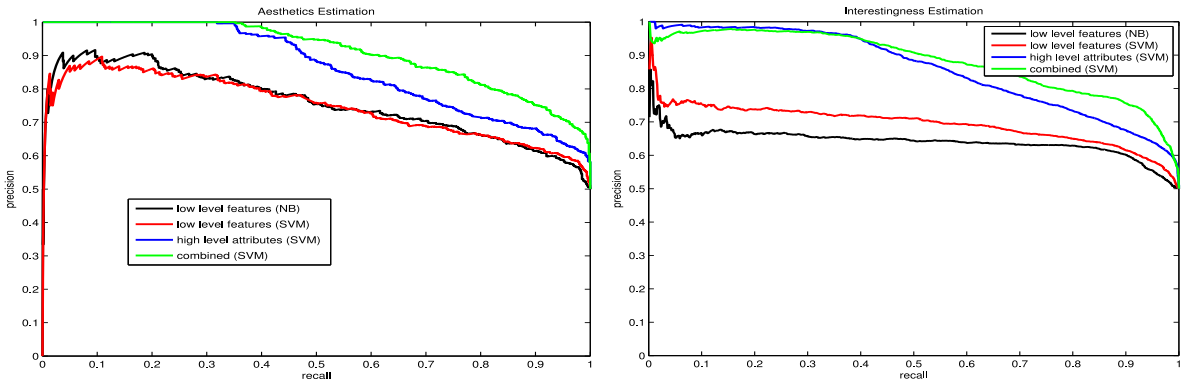

Figure 5. Precision-Recall curves for aesthetics estimation (left) and interestingness estimation (right - averaged over 6 queries and general Flickr set). In both cases we show estimation using low level image features with Naive Bayes classification (ie method proposed in *Ke et al [11]* – **black**), the same low level image features using SVM classification (**red**), *our high level describable attributes* with SVM classification (**blue**), and a *combination* of low level features and high level attributes with SVM classification (**green**). For both aesthetics and interestingness our high level attributes (blue) produce significantly powerful classifiers than the previous method (black), and can provide complimentary information when used in combination with low level features (green).

to form a single visual dictionary which is used for all of the content attribute types.

For each of these attributes we use an appropriate data set for training and testing. For "presence of animals" this is the Animals on the Web dataset [2] with images from all 10 animal categories merged into an animal superclass. For "indoor-outdoor" this is 2000 images collected from Flickr (half indoor, half out). For the 15 "scene type" attributes this is the 15 scene category dataset [19, 6]. Precision-recall curves for each attribute (subsampled for scenes for clarity of presentation) are shown in fig 4. Though it is well known that recognition of specific animal categories is very challenging [2], we do quite well at predicting whether *some* animal is present in an image. The indoor-outdoor classifier is very accurate for most images. For scenes, natural scene types tend to be more accurate than indoor scenes.

2.3. Sky-Illumination Attributes

Lighting can greatly effect perception of an image – *e.g.* interesting conditions such as indirect lighting can be more aesthetically pleasing. Because good indoor illumination is still a challenging open research problem, we focus on natural outdoor illumination through 3 attributes:

- Clear skies – photos taken in sunny clear conditions.
- Cloudy skies – photos taken in cloudy conditions.
- Sunset skies – photos taken with sun low in the sky.

To train our sky attribute classifiers we first extract rough sky regions from images using Hoeim et al's work on geometric context [9]. This work automatically divides image regions into sky, horizontal, and vertical geometric classes using adaboost on a variety of low level image features. On the predicted sky regions we compute 3d color histograms in HSV color space, with 10 bins per channel, and train 3 sky attribute SVMs using 1000 manually labeled images from Flickr. The classifiers produced are extremely effective (99%, 91.5% and 96.7% respectively).

3. Estimating Aesthetics & Interestingness

3.1. Aesthetics

The first task we evaluate is estimating the aesthetic quality of an image. Here the goal is to differentiate between images of high photographic quality from images of (low) snapshot quality.

Experiments: Because aesthetic quality is by nature subjective, we make use of human evaluated images for training and testing. We collect a dataset of 16,000 images from the DPChallenge website[1]. These images have been quantitatively rated by a large set of human participants (many of whom are photographers). We label the top 10% rated photos as high aesthetic quality, and the bottom 10% as low quality to allow a direct comparison to Ke et al [11]. Going further down in the ratings is possible, but increases ambiguity in ratings. Half of each of these sets is used for training, while the remaining half is used for evaluation.

To estimate aesthetic quality we train an SVM classifier where the input image representation is the outputs of our 26 high level describable attribute classifiers (fig 2 shows our pipeline). For comparison we also reimplement the baseline aesthetics classifier used in Ke et al [11]. We show results of their original Naive Bayes classification method (fig 5, left plot black) and also train an SVM on their low level features (fig 5, left plot red). Our high level attributes produce a significantly more accurate ranking than the previous approach, and when used in combination with these low level features can produce an even stronger classifier (fig 5 left plot, green). This suggests that our high level

[1] http://www.dpchallenge.com/

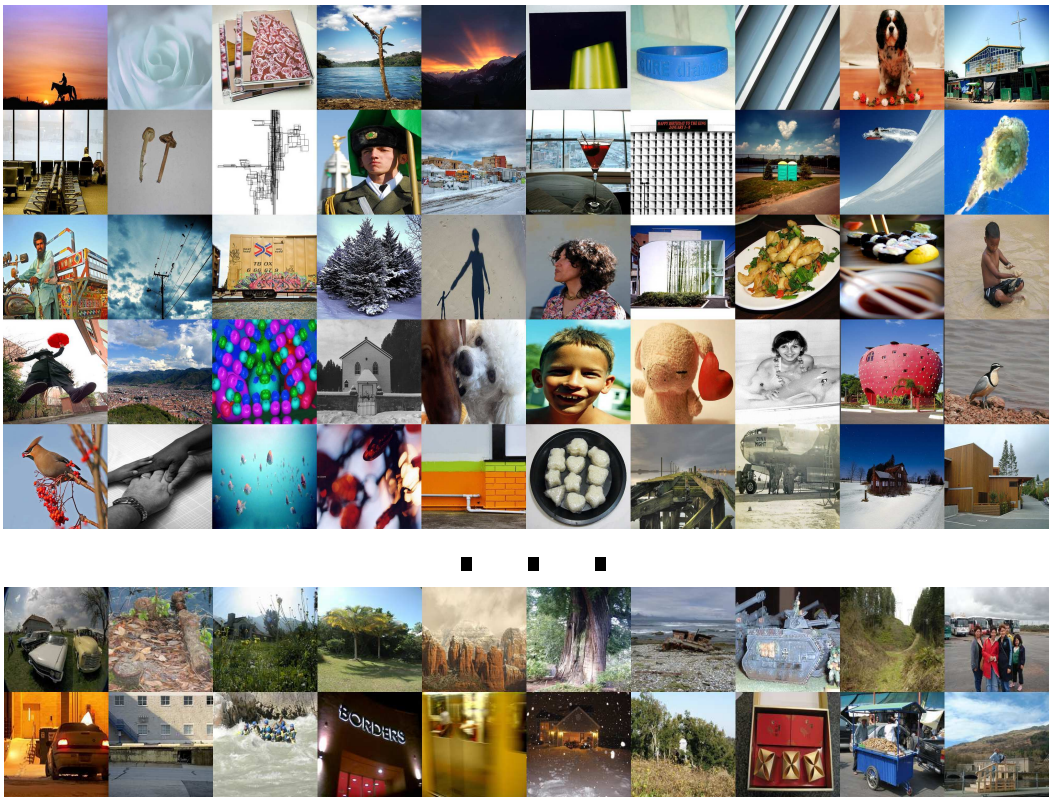

Figure 6. General Flickr photos ranked by interestingness. Top 5 rows show the first 50 images ranked by our interestingness classifier. Bottom two rows show the last 20 images ranked by our interestingness classifier.

attributes are providing a source of useful complimentary information to purely feature based approaches.

3.2. General interestingness

We also apply our describable attributes to a related, but deceptively different problem of estimating interestingness of photos. While DPChallenge directly measures aesthetic quality through user ratings, Flickr's "interestingness" measure[2] is computed more indirectly through analysis of social interactions with that photo (viewing patterns, popularity of the content owner, favoriting behavior, etc).

Experiments: For our general interestingness task we collect a dataset of 40,000 images from Flickr using interestingness-enabled Flickr searches on time limited queries. The top 10% of these images are used as positive examples for our interestingness classifier, while the bottom 10% are used as negative examples (splitting this set in half for training and testing).

Again here we train an SVM classifier to predict interestingness using our 26 describable attribute classifications as input (fig 5, right plot blue). For comparison, we also train an interestingness classifier on the low level features used in Ke et al [11] using their original Naive Bayes approach (fig 5, right plot black), and using an SVM classifier (fig 5, right plot red). Lastly we train a combined classifier on their low level features and our high level attribute classifications, a 32 dimension input feature vector (fig 5, right plot green). In fig 6 we show images ranked by automatically predicted interestingness score. The top 5 rows show 50 highly ranked images, and the bottom 2 rows show 20 low ranked images and reflect the variation in interestingness between the top and bottom of our ranking.

In fact, our method performs extremely well at estimating interestingness (fig 5 right plot). The high level attributes produce a powerful classifier for predicting interestingness (fig 5 blue), and improve somewhat with the addition of low level features (fig 5 green). Compared to aesthetics classification, our interestingness classifier shows an even larger increase in performance over the previous method (fig 5, black vs blue).

3.3. Query specific interestingness

Lastly, we introduce a method to produce query specific interestingness classifiers. In general we expect some of our attributes to be more useful for predicting interestingness than others. We also expect that the usefulness of an attribute might vary according to the specific search query used to collect images – *e.g.* low DoF may be more useful for predicting interestingness of images returned for the

[2] http://www.flickr.com/explore/interesting/, patent #20060242139

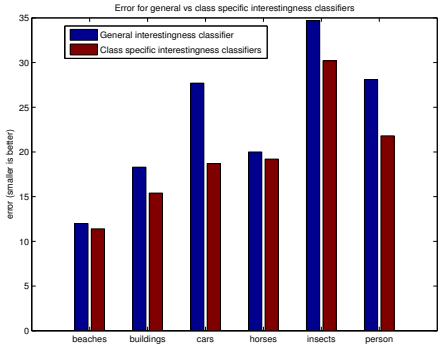

Figure 7. Query specific error rates for interestingness prediction. For some categories, the query specific classifiers (blue) has significantly lower error than the general interestingness classifier (red).

query "insect" than for the query "beach".

Experiments: We collect a dataset of images from Flickr using 6 different query terms: "beach", "building", "car", "horse", "insect", and "person", retrieving 20,000 images for each query ranked by interestingness. Again the top 10% are labeled as positive, bottom 10% as negative and the collection is split in half for training and testing. For each query collection we train an interestingness predictor. We then evaluate the accuracy of our general interestingness classifier vs using our query specific classifiers to rank images from the held out test set. For some queries, the query specific classifiers outperform the general interestingness classifier (*error rates* are shown in fig 7).

Ranked results for some of our query specific interestingness classifiers are shown in fig 8 where top 3 rows show 30 most highly ranked images and bottom rows show the 10 lowest ranked images for each query. At the top of the "beach" ranking we observe very beautiful, clear depictions, often with pleasant sky illuminations. At the bottom of the ranking we see more cluttered images often displaying groups of people. For the insect query, the top of the ranking shows images where the insect is the main subject of the photograph, and a low DoF is often utilized for emphasis. In general for each category the top of our ranking shows more picturesque depictions while the bottom shows less clean or attractive depictions.

Acknowledgments

This work was supported in part by NSF Faculty Early Career Development (CAREER) Award #1054133.

References

[1] O. Axelsson. Towards a psychology of photography: dimensions underlying aesthetic appeal of photographs. In *Perceptual and Motor Skills*, 2007. 1659

[2] T. L. Berg and D. A. Forsyth. Animals on the web. In *CVPR*, 2006. 1661

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006. 1658, 1660

[4] R. Datta, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, 2008. 1658

[5] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1658

[6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 1661

[7] V. Ferrari and A. Zisserman. Learning visual attributes. *NIPS*, 2007. 1658

[8] J. Gardner, C. Nothelfer, and S. Palmer. Exploring aesthetic principles of spatial composition through stock photography. In *VSS*, 2008. 1659

[9] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005. 1661

[10] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001. 1659

[11] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006. 1658, 1661, 1662

[12] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1657, 1658

[13] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1658

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching. In *CVPR*, June 2006. 1660

[15] D. G. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 2004. 1660

[16] T. Lui, J. Sun, N.-K. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR*, 2007. 1659

[17] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008. 1658

[18] C. Nothelfer, K. B. Schloss, and S. E. Palmer. The role of spatial composition in preference for color pairs. In *VSS*, 2009. 1659, 1660

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001. 1661

[20] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In *Attention and Performance*, 1981. 1658

[21] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *CVPR*, 2005. 1659

[22] K. B. Schloss and S. E. Palmer. Color preferences. In *VSS*, 2007. 1659, 1660

[23] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computational visual attention model. In *ACM MM*, 2009. 1658

[24] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *PCM*, 2004. 1658

[25] H. Tong, M. Li, H. Zhang, C. Zhang, J. He, and W.-Y. Ma. Learning no-reference quality metric by examples. In *ICMM*, 2005. 1658

[26] P. Viola and M. Jones. Robust real-time object detection. In *IJCV*, 2001. 1660

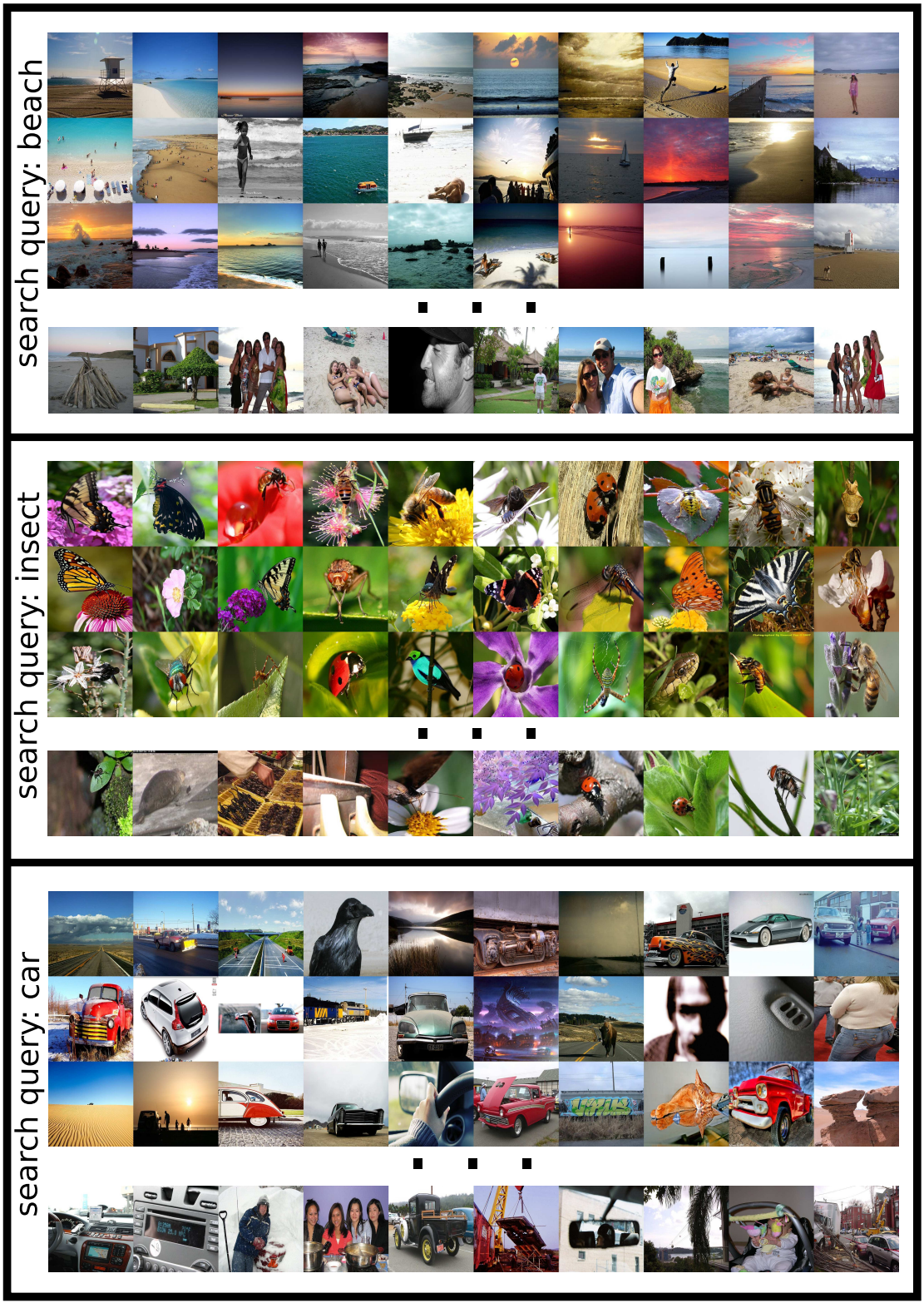

Figure 8. Query specific interestingness ranking for search terms (beach, insect, car). Top three rows for each query show the most highly ranked images. Bottom rows for each query show the least highly ranked images.