Leveraging expert feature knowledge for predicting image aesthetics

Michal Kucer, *Student Member, IEEE,* Alexander C. Loui, *Fellow, IEEE,* and David W. Messinger, *Member, IEEE*

Abstract—The ability to rank images based on their appearance finds many real-world applications such as image retrieval or image album creation. Despite the recent dominance of deep learning methods in computer vision which often result in superior performance, they are not always the methods of choice because they lack interpretability. In this work, we investigate the possibility of improving image aesthetic inference of convolutional neural networks with hand-designed features that rely on domain expertise in various fields. We perform a comparison of hand-crafted feature sets in their ability to predict fine-grained aesthetics scores on two image aesthetics datasets. We observe that even feature sets published earlier are able to compete with more recently published algorithms and, by combining the algorithms together, one can obtain a significant improvement in predicting image aesthetics. By using a tree-based learner, we perform feature elimination to understand the best performing features overall and across different image categories. Only roughly 15 % and 8 % of the features are needed to achieve full performance in predicting a fine-grained aesthetic score and binary classification respectively. By combining hand-crafted features with meta-features that predict the quality of an image based on CNN features, the model performs better than a baseline VGG16 model. One can, however, achieve more significant improvement in both aesthetics score prediction and binary classification by fusing the hand-crafted features and the penultimate layer activations. Our experiments indicate an improvement up to 2.2 % achieving current state-of-the-art binary classification accuracy on the AVA dataset when the hand-designed features are fused with activation from VGG16 and ResNet50 networks.

Index Terms—computational aesthetics, image aesthetics, hand-crafted,aesthetic quality assessment

1 Introduction

With continuous miniaturization of silicon technology and proliferation of consumer and cell-phone cameras, we have seen an exponential increase in the number of images that are captured [1]. Whether the images are stored on personal computers or reside on social networks (e.g. Instagram, Flickr), the sheer number of images calls for methods to determine various image properties, such as object presence or appeal, for the purpose of automatic image management and curation. One of the central problems in consumer photography centers around determining the aesthetic appeal of an image.

The problem of determining the aesthetic appeal of an image is challenging because the overall aesthetic value of an image is dependent on its technical quality, composition, emotional value, etc. In determining the aesthetic value of an image, the algorithms follow a similar pipeline to other branches of Computer Vision, such as object detection: a set of image features is extracted from an image which is then used as an input to a classifier or regressor for further processing.

Many of the early algorithms for inference of image aesthetics relied on carefully chosen and crafted features based on expert knowledge, e.g. established photographic rules [2], [3] such as those seen in Figure 1. These, however, went out of favor and were replaced by generic features based on various local descriptors and convolutional neural networks. Although these networks are superior in their capacity to learn high-level semantic information from low-level pixel information, it is possible that the networks may not discover some essential knowledge, e.g. global texture information contained in the gray-level co-occurrence matrix, even when appropriate optimization is in place.

(a) Rule of Thirds (b) Symmetry

Fig. 1: Common photographic rules used in capturing aesthetically pleasing photographs.

In this paper we conduct a comprehensive study of hand-designed features that rely on expert knowledge from various fields, and we explore to what extent can hand-crafted features aid learning-based features in predicting image aesthetics. The major contributions of this paper are listed as follows:

- We analyze and compare a wide variate of hand-crafted features in their ability of predicting continuous and binary aesthetic scores.

- *M. Kucer and D. W. Messinger are with the Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, 14623. E-mail: mxk7721@rit.edu*
- *Alexander C. Loui is with Kodak Alaris Inc., Rochester, NY, 14615.*

Manuscript received September, 2017; revised April 22, 2018.

- We perform feature elimination for various tasks (classification, regression, categories) to uncover the best performing features for them.
- We investigate the possibility of fusing hand-crafted features with learned features for improving aesthetic inference.

The remainder of the paper is organized as follows. In Section 2. we summarizes the related work and datasets used in this paper. Section 3. presents the analysis of the ability of hand-designed features in predicting aesthetics, and performs feature elimination to surface the best performing features. Section 4. explores the possibility of fusing hand-crafted features with learned features from convolutional neural networks. Concluding remarks and suggested future work are discussed in Section 5.

2 Related Work

Computational Aesthetics is a sub-discipline of Computer Vision interested in developing algorithms that predict the aesthetics value of an image. The features that the algorithms compute can be described to be either hand-crafted, generic, or learned (deep learning features).

Hand-crafted features

Early algorithms in image aesthetics looked for an inspiration in fields of image processing, psychology, perception, and photography. These algorithms create features that approximate the various guidelines from professional photography, such as the rule of thirds or approximate the quality of an image by estimating the blur present. Early works include Datta et al. [4] and Ke et al. [5], where they attempt to predict whether the images are high/low quality and professional photographs/snapshots respectively. Luo et al. [6] and Wong et al. [7] recognize the importance of the subject in a photograph. Using a clarity-contrast based and saliency based subject extraction method respectively to extract the subject regions, they compute an image descriptor that combines global features on the whole image and subject/background specific features. Dhar et al. [8] use low-level image features to estimate high-level image attributes (e.g. presence of specific objects in the image) and use these features along with low-level image attributes to describe the image. Tang et al. [9] consider the content of images by using different subject extraction methods based on image category and computing additional features on an image with faces. Redi et al. [10] study the possibility of uncovering beautiful images in a set of images that are of low popularity.

Generic features

Several approaches use generic descriptors of an image, such as the scale-invariant feature transform (SIFT) descriptors or local patches of colors, to create a bag-of-words model to describe images. Nishiyama et al. [11] extract local color patches from a regular grid and create a bag-of-color-patches descriptor for each image. Marchesotti et al. [12] extract SIFT and local color descriptors and construct a Fisher Vector descriptor for each image.

Learned features

Beginning with the strong performance of Krizhevsky et al. [13] in the ImageNet Challenge, Deep Learning techniques have started to dominate the research in many areas. The domain of aesthetic inference is no different as can be seen in the recent review by Deng et al. [14], as most of the top algorithms for various tasks and datasets are held by deep learning approaches.

Lu et al. [17] are among first to tackle the problem of aesthetics inference that solely uses Convolutional Neural Networks (CNN). They conduct a thorough study of several network architectures and experiment with constructing multiple column networks with varying inputs. As different photographic rules (e.g. rule of thirds or color harmony) consider properties on different scales of the image, authors used both the global and the local view (smaller random crop) of the image to train the networks. To get around the problem of varying image sizes, the authors experiment with various image transformations to fix the size of the image, e.g. center-crop. Since the global and local details of the image are important, the authors propose a Double Column CNN (DCNN, where one of the columns accepts a global-view of the image and the other a local-view). It is essentially a network consisting of two separate Single Column CNN (SCNN) whose outputs are in the end combined to produce a single score. Because of the large intra-class variation in aesthetics scores, Lu et al. propose to use higher-level style labels that are present for a subset of images present in the AVA dataset as additional features. They train an additional Style SCNN to recognize the various style labels and use the output to augment the features computed by the DCNN. The Style-SCNN is trained to recognize different style attributes (e.g. complementary colors, motion blur or the rule of thirds). The trained network is then used to extract the Style features for the rest of the images in the AVA dataset. These features are then concatenated with the features computed by the DCNN network and used to determine the final aesthetics score. A very interesting detail arises when looking at the images correctly classified by the DCNN and incorrectly by SCNN.

Wang et al. [18] take inspiration from Neuro-aesthetics and Neuroscience of Vision to propose a novel architecture that aims to tackle the problem of binary classification and the distribution of aesthetics scores. They propose a model called the Brain-Inspired Deep Network (BDN) and is primarily composed of two parts: a Parallel Pathways layer and a High-level Synthesis Network. The Parallel Pathways layer is inspired by the parallel pathway processing of the human cortex, which decomposes the visual scene into several representations that encode information such as intensity and edge information in the image. In this layer they convert the RGB data into the HSV format and use each H, S, V as one of the parallel representations for the image. As was shown previously in [8], high-level attributes are successful in augmenting aesthetics prediction if used as mid-level features. Therefore Wang et al. decided to train fourteen fully-convolutional networks (FCN) trained in a supervised manner to predict the fourteen binary style labels available with the AVA dataset. The activations of the mid-level convolutional networks for each of the fourteen

Types of Features	Feature	Description
High-Level Features	Face Detection [9]	Using a detector to uncover the presence of faces in the image
	Face Shadow [9]	Approximating the quality of lighting on faces.
	Average Region Saliency [10]	Measures salience of objects in different parts of the image
Affect	Affective Dimensions: Pleasure, Arousal, Dominance	Indicators of emotion calculated by combining average Saturation and Brightness as defined by [15]
	HSV Statistics [4], [10]	Measure of the mean and spread of the HSV image channels.
Aesthetics	Rule of Thirds [4], [9], [10]	Guideline in photography for placing the subject within the image.
	Depth of Field [4]	How well is the background separated from the foreground?
	Colorfulness [4]	How different is the color distribution in an image from an ideal one.
	Color Harmony [16]	Features approximating how pleasing are different color combinations.
Texture	GLCM Entropy & Skewness	Measures of texture based on Gray Level Co-occurrence Matrix
	Wavelet-based Texture [4]	Measures of spatial smoothness based on Daubechies wavelets

TABLE 1: Type, name and description of the variety of features that the algorithm considers

style label are used in parallel as features for the high-level synthesis network (thus virtually decomposing image into several representations encoding different information).

Guo et al. [19], similarly to previous CNN work, proposes to use parallel Deep CNN (PDCNN) architecture. They utilize an architecture similar to the AlexNet [13]. Since CNNs are prone to both over and under-fitting, they propose to use PDCNN to control the complexity of the system by stacking n columns in parallel - n-PDCNN. They show that the performance increases by combining up to three parallel networks, while it drops when adding more networks.

Kong et al. [20] proposes AlexNet inspired architecture to predict various image attributes. First, they create simple regression network to predict aesthetic rating by minimizing the Euclidean Loss. Subsequently, they adopt a Siamese Network architecture [21] to jointly optimize the network to both predict an aesthetic score as well as a relative ranking of the two images. Similar to [18], Kong et al. predict the aesthetic attributes of images augmenting the network with an auxiliary task of predicting attributes from the same activations that are used to predict the aesthetic score. Lastly, the network is used to predict image categories.

Datasets

In the process of computing features and evaluating the algorithms, the following datasets are used: Aesthetic Visual Analysis (AVA) [22], CUHKPQ [9], HiddenBeauty [10], and a Kodak Aesthetics dataset [23].

The *CUHKPQ* dataset contains more than 17,690 images divided into seven semantic categories with binary labels indicating high or low quality. In order to assign labels to the images, each image was viewed by ten people who labeled the image as High or Low-quality. An image was kept and assigned a final label if at least 8 out of 10 people agreed with their assessment of the image. We primarily use the CUHKPQ in the computation of image features that required reference high/low quality data. For example, Ke et al. [5], one of the algorithms considered here, computes a color distribution feature, which calculates the number of high-quality photos retrieved by the nearest neighbor search.

The *Hidden Beauty of Flickr Pictures* (HiddenBeauty) dataset was collected as part of an effort to surface the "hidden gems" among the pictures that have very low popularity/interestingness as measured on Flickr [10]. More than 15,000 images were chosen from the sample of nine million images from the larger YFCC100M dataset [24]. Although the HiddenBeauty database is not the largest database, we chose to use it because as image was rated on a five-point scale based on metrics that were clearly described to each rater, and thus it attempts to minimize the bias of the notion of what is "high quality". The labels were collected via the CrowdFlower crowdsourcing platform. Each image was labeled by at least five different people (each evaluator having a top track record on the platform). Each image belongs to one of four categories - human, nature, urban, people - and its aesthetic score is the mean rating of all of the scores.

The Kodak Aesthetics dataset is an extended version of the dataset described in Jiang et al. [23]. The dataset was created to resemble the variety of images found in consumer photography. It consists of more than 1,500 images each rated by four people on the 1-100 scale. The ground truth score for an image is the average of its four ratings.

The *Aesthetic Visual Analysis* (AVA) dataset is one of the largest datasets available for working with aesthetic preferences [22]. The images were sourced from *www.dpchallenge.com*, a website housing a community of amateur and professional photographers. Each image in the dataset received between 78 and 549, with an average of 210 votes per images. Each image is given a score on scale $1-10$ and then the average rating is considered to be the ground truth aesthetic score for the image. Along with aesthetic ratings, images come with 66 semantic and 14 photographic style annotations (e.g. High Dynamic Range, Soft Focus, etc.).

3 Aesthetic Assessment with Hand-crafted features

To better understand the utility of hand-crafted features in aesthetic assessment, this section compares the performance of a selection of algorithms and investigates an approach for selecting a subset of features. For our investigation, we selected algorithms enabling a wide variety of image features to be considered, e.g. different measures of photo quality such as image blur [4], [5], [25], image composition and content [6], [9] and generic features [12], [25]. Table 1 shows and describes a selection of features considered. For specific details on the individual features in each feature set, please see the selected references. The features selected for comparison originate from the following feature sets:

1) Datta et al. (DATTA) [4]

	DATTA	KE	PQ	CBPQ	FV	YCF-HC	YCF	EPFL	VQ	YHB	NoDLFV	All	CNN
HB	0.413	0.431	0.296	0.458	0.258	0.478	0.540	0.421	0.434	0.453	0.568	0.600	0.589
Kodak	0.571	0.547	0.293	0.297	0.310	0.589	0.67	0.384	0.509	0.548	N/A	0.733	0.636

TABLE 2: Comparison of the algorithm performance in predicting aesthetics score in terms of the correlation coefficients for the HiddenBeauty and Kodak datasets.

2) Ke et al. (KE) [5]
3) Subject-based Photo Quality (PQ) [6]
4) Content-based Photo Quality (CBPQ) [9]
5) Aesthetics using Generic Image features (FV) [12]
6) Yahoo Complete Framework for Image Aesthetics (YCF) [25]
7) EPFL Context Image Aesthetics (Global features) (EPFL) [26]
8) Video Aesthetics (VQ) [16]
9) Yahoo HiddenBeauty Algorithm (YHB) [10]

Both pre-trained CNN features and the various generic features (e.g. SIFT) are either used independently to predict the quality of images or to create higher-level meta-features. In total, we extract a total of 331 numerical features: 54 for DATTA, 10 for KE, 10 for PQ, 16 for CBPQ (20 for images with humans), 27 for YCF, 14 for EPFL, 149 for VQ, and 47 for YHB. Additionally, we extract the Fisher Vector descriptors as described in [12] and introduced in [27]. In implementing the DATTA algorithm we avoid the features labeled $f8$ and $f9$ in the paper, because the computation of image uniqueness (computed as the mean distance to the top 20 and 100 closest matches) was tied to a selection of 1000 images which were unavailable. Additionally, for the computation of the human-related features in CBPQ, we use the deep learning based face detector in the *dlib* machine learning library [28], which is more accurate in terms of detection performance and False Positive retrieval.

3.1 Learning framework

Often one of the most important aspects of building predictive systems is the selection of a suitable learning framework, one that achieves good generalization performance and speed of evaluation. Earlier algorithms use a variety of techniques to predict the aesthetic image descriptor, e.g SVM, Neural Networks or Random Forest. To understand the differences and performance of individual algorithms and features, the same learning framework is used on top of each feature set. The model / learner in use is an ensemble learning technique known as Gradient Boosted Trees as it achieves excellent generalization performance supported by both theory and practice [29]. An advantage of using a Boosted Tree learner is its ability to simultaneously quantify the importance of features and train a model. The notion of feature importance is measured by three metrics: gain, cover, and frequency [30].

3.2 Methodology

To evaluate the performance of the chosen algorithms and individual features, a separate model is trained for each algorithm. In order to maximize the utility of the datasets, we evaluate the performance of the algorithms using k-Fold Cross-validation (CV). In k-Fold CV, the dataset $\mathcal{D}$ is split into k non-overlapping folds/parts $\mathcal{D}_1, \ldots, \mathcal{D}_k$ (we chose $k = 10$ based on [31]). To estimate the performance metric on a dataset, the learner (or inducer $\mathcal{I}$ as in [31]) is trained and tested k times. Given a fold $i \in \{1, \ldots, k\}$, the model is then trained on the dataset $\mathcal{D}\backslash\mathcal{D}_i$ and tested on $\mathcal{D}_i$. Thus we iterate through each fold, treating it as the test set and the other $k-1$ folds as the training set for our learner, allowing us to utilize each image in both training and testing. The cross-validation estimate of accuracy is given by calculating the accuracy / correlation coefficient for the concatenated vectors from all folds [31] as

$$acc_{CV} = \frac{1}{m} \sum_{\langle x_i, y_i \rangle} \delta\left(\mathcal{I}\left(\mathcal{D}\backslash\mathcal{D}_i, x_i\right), y_i\right), \tag{1}$$

where m is the total number of samples in the dataset, $\mathcal{I}\left(\mathcal{D}\backslash\mathcal{D}_i, x_i\right)$ is $\hat{y}_i$, the label given to x_i that was trained on the dataset consisting of all of the folds except one that includes x_i. To prevent bias towards any algorithm or feature set, the training/prediction of the aesthetics scores was performed across all tests with the same parameters[1] (other learner parameters are set at their default values). In training the models to predict the continuous aesthetic score, we optimize the mean square error

$$l_\theta = \sum_i \left(y_i - \hat{y}_i\right)^2, \tag{2}$$

and for predicting the binary High / Low score we optimize the logistic loss

$$l_\theta = \sum_i \left[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)\right], \tag{3}$$

where y_i is the ground truth aesthetics score, and $\hat{y}_i$ is the predicted aesthetics score, and θ are parameters optimized to achieve the best performance for a given loss function. We quantify the performance of each trained model by either calculating the average accuracy for binary labels as defined in Equation 1 or calculating the predicted aesthetic score for each fold and then calculating the correlation coefficient between the predicted and ground truth values for each image.

3.3 Comparing different algorithms

First, we compare the different algorithms to each other in terms of their ability to predict the continuous aesthetics scores of the HiddenBeauty and Kodak datasets. The YCF algorithm contains a feature which predicts the probability of an image being high quality based on the deep learning features extracted from the second to last layer of a CNN (ImageNet pre-trained VGG16 model [32]). Therefore, in addition to considering the nine algorithms detailed above, we

1. The particular values for the XGBoost Tree Learner [29] are as follows: max_depth, n_estimators, subsample, colsample_bytree, and colsample_bylevel are respectively set at 8, 100, 0.9, 0.9, and 0.9

Regression		Classification	
ALL	NoDLFV	ALL	NoDLFV
A6 f25: DL Probability	A2 f5: Blur	A6 f25: DL Probability	A2 f5: Blur
A2 f5: Blur	A8 f92: White Balance 6	A2 f5: Blur	A8 f92: White Balance 6
A6 f26: Sift FV Probability	A1 f55: V DOF Ind	A1 f54: S DOF Ind	A9 f17: Itten H07
A2 f9: EdgeDist_Low	A4 f15: Spatial Complexity Rela	A9 f17: Itten H07	A9 f27: Itten S05
A8 f92: White Balance 6	A2 f9: EdgeDist_Low	A6 f26: Sift FV Probability	A2 f9: EdgeDist_Low
A4 f9: Dark Channel	A1 f22: Size feat	A2 f7: Tong_BlurExtent	A2 f1: BBox_Edges
A1 f1: mean intensity	A2 f1: BBox_Edges	A2 f9: EdgeDist_Low	A1 f22: Size feat
A4 f15: Spatial Complexity Rela	A2 f7: Tong_BlurExtent	A9 f13: Itten H03	A9 f14: Itten H04
A1 f5: ROT mean H	A4 f9: Dark Channel	A8 f92: White Balance 6	A4 f14: Spatial Complexity Bkgd
A2 f7: Tong_BlurExtent	A1 f7: ROT mean V	A4 f9: Dark Channel	A1 f55: V DOF Ind

TABLE 3: Top 10 performing features for regression / classification on ALL / DLFV features sets.

train three additional models: one which only considers the scores predicted by a model trained on CNN features, one with only the hand-crafted features from YCF, and a model which considers all but features that predict the probability of being High-Quality image based on the CNN features, and SIFT and Color Fisher Vectors (denoted "NoDLFV" in Table 2). Table 2 shows the model performance for the various feature sets as evaluated by the correlation coefficient for each model.

As we can see from Table 2, even features crafted by early algorithms are effective for predicting photo quality, as evidenced by the competitive correlation coefficient we can see for the first two algorithms in 2006 (DATTA, KE) as compared to the more recent methods (VQ, YHB). If we compare the performance of the algorithms between the two datasets, we see that although the algorithms perform better on the Kodak dataset, the algorithms generally exhibit the same trend in the performance on both datasets. The better performance of algorithms on the Kodak dataset can be explained by the way scores were obtained: in the Kodak dataset, each image was scored by the same four people, as opposed to the CrowdFlower platform where a diverse group of people rate each image (resulting in a larger variance in the scoring from image to image).

The last column of Table 2 shows the results for the pre-trained CNN features, obtained as described above from the ImageNet pre-trained VGG16 model. We can see that, despite the absence of fine-tuning the features perform very well, giving us the second/third best results among all of the feature sets for the HiddenBeauty / Kodak dataset respectively.

One of our original goals was to observe ways of combining hand-crafted features and deep learning, and thus we observe two models: "NoDLFV" and "ALL", which will be described in the next section. Combining all hand-crafted features results in an improvement as compared to the YHB (the best single algorithm). Furthermore, one of the ways we can combine HC features with DL, is to use pre-trained CNN features to compute a Quality meta-feature, and then concatenate it to them, resulting in improvement in performance of predicting aesthetics as can be seen in Table 2.

3.4 Feature Elimination

Although combining all of the features improves results by a small margin, computing the features for all algorithms is computationally inefficient. Therefore in this section we investigate how many of the features are actually needed to achieve a good performance in predicting aesthetics on the HB dataset. To determine the top features, we will perform *Recursive Feature Elimination* (RFE) [33], which itself is an instance of Backward Feature Elimination [34]. RFE is an iterative procedure and can be simply described as the following sequence of steps [33]:

1) Train the classifier (optimizing the parameters θ with respect to a loss function l_θ)
2) Compute the ranking criterion $\mathcal{D}(X_i)$ for each feature X_i
3) Remove the feature with the smallest ranking criterion.

At each step we train a learner to predict the aesthetics scores and based on ranking of all features , we remove the feature with the lowest ranking criterion, and retrain the model with remaining features. We use the gain of each feature (defined as "improvement in accuracy brought by a feature to the branches it is on" [30]) as a ranking criterion for the individual features as it is the most intuitive way to measure feature importance among the three metrics. The gain-based ranking criterion is similar to the Mean Decrease Impurity importance (MDI) [35]

$$\mathcal{D}(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T: v(s-t)=X_m} p(t) \Delta i(s_t, t) \qquad (4)$$

where $i(t)$ is any impurity measure (in our case the impurity or gain of each feature will depend of the loss function used to optimize the learner). In order to observe how the performance of the model changes with each removed feature, at each step we perform a 10-fold cross-validation on the current features where we predict the mean aesthetic score. For a more formal description of RFE, please see Section 3.2 of [33].

Aesthetic inference

Figure 2 shows the variation of r^2 in predicting the mean aesthetic score with respect to the number of considered features (the abscissa covers a shorter range, since there is no change in performance for more than 100 features). As we can see in Figure 2, many of the features contain information that could be considered complimentary and, thus, do not improve the performance past roughly 40 features. As can be expected, "ALL" features perform slightly better in terms of predicting the overall score than the "NoDLFV". Although the performance of the "NoDLFV" features is

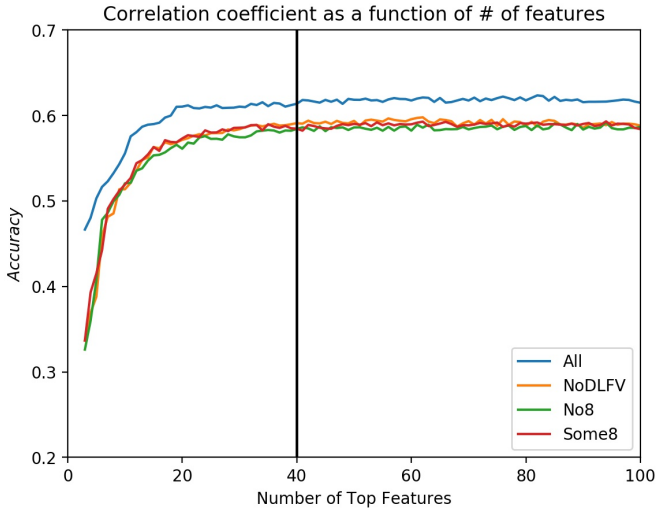


Fig. 2: Regression performance as the function of top k features on the HB dataset. The vertical line at $k = 75$ indicated a point after which the regression performance remained approximately constant.

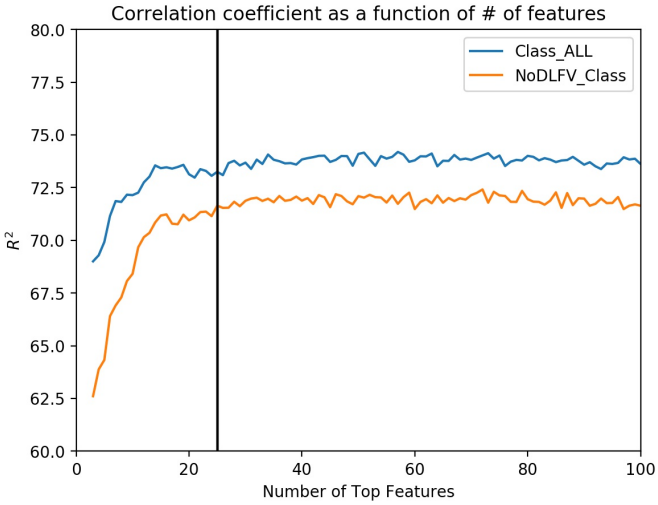


Fig. 3: Classification performance as the function of the number of top k features HB dataset. The vertical line at $k = 25$ indicated a point after which the classification performance remained approximately constant

arguably constant even with 250 features removed, we see a much sharper drop in the performance as compared to "ALL" if we keep removing additional features past this point. This could indicate the strength of the DL Probability feature and how much the "quality" of the image is related to the aesthetics score.

Figure 2 shows "Some8" and "No8" feature sets in which part or all of the features from VQ algorithm were respectively removed in order to study its impact on the regression performance since when observing the top 75 features, a large proportion of the features came from this algorithm. As can be seen in the performance of the learner is the same and thus corroborating the notion of the complementarity/redundancy of some of the features. This can further be seen by comparing the top features of the different features sets. In NoDLFV, many of the features from VQ were pertaining to Color Harmony and Colorfulness. Once all of the features from the VQ algorithm were removed, features describing similar information took their place, e.g. the Color Harmony and Hue Complexity features from CBPQ [9].

Binary Classification

In order to perform classification, the HiddenBeauty dataset is split at the mean score μ and assign images with scores $\geq \mu + \delta$ to the "high" quality class and images with scores $\leq \mu - \delta$ to the "low" quality class. Then we compute the binary accuracy, as defined in (1), of the predicted labels for the HiddenBeauty dataset based on the 10-fold CV, as described previously in section 3.2. Figure 3 shows the classification performance as a function of the top k features. Similarly to regression, the "ALL" features achieve a better performance, because of the DL probability feature as can be seen from Table 3. It was found that, for classification, only 25 features are needed to achieve the full performance as opposed to the top 40 for regression. This can be attributed to the complexity of predicting the continuous aesthetic scores: trying to predict the aesthetics scores is much harder task, since notion of beauty or aesthetics of an image is a very subjective and thus we are likely to be some learning of the underlying noise. Dividing the images into High or Low Quality and predicting binary label is an easier task. Table 3 lists the top ten features (as measured by the gain factor described above) as an example for the different tasks and feature sets tested for classification and regression, and interestingly the top two performing features do not change from regression to binary binary classification. In all tasks and datasets, features that measure technical quality of images (Blur or White Balance) rank very high in importance, however in classification we see much higher importance placed on features related to color (Itten).

3.5 Model and feature analysis by categories

The content of images is known to affect the attributes that describe the image the best [9]. In our analyses of the category-specific features, we consider the HiddenBeauty dataset since it separates all of its images into one of following categories based on content: people, urban, nature, and animals. For the purpose of this investigation, we assume that category of an image is known to us beforehand. We measure the importance of all the features by first concatenating the feature extractions from the explored feature sets, and using previously described RFE to uncover the top performing features for each category.

The resulting r^2 values computed with 10-fold CV for each category are listed in Table 5. As we can see, images of animals achieve the best performance in predicting their aesthetic score, followed by the urban, people and nature categories. The nature category proves particular difficult due to the diversity of its contents: the nature category includes a wide variety of images depicting landscapes, plants and animals (i.e. smaller animals like bees on a flower or a bird flying around tree). Figure 4 shows the regression performance as a function of the number of

Animals	Nature	People	Urban
A6 f25: DL Probability	A6 f25: DL Probability	A6 f25: DL Probability	A6 f25: DL Probability
A2 f1: BBox_Edges	A2 f5: Blur	A2 f5: Blur	A9 f30: Itten V03
A2 f9: EdgeDist_Low	A9 f27: Itten S05	A2 f9: EdgeDist_Low	A4 f16: Hue Complexity
A2 f5: Blur	A9 f34: Symmetry	A2 f7: Tong_BlurExtent	A8 f1: Sal Region Area
A4 f9: Dark Channel	A1 f7: ROT mean V	A8 f92: White Balance 6	A2 f5: Blur
A8 f10: Dark Channel 5	A1 f4: mean hue	A1 f6: ROT mean S	A6 f21: Noisiness
A7 f13: Sharpness	A6 f26: Sift FV Probability	A6 f9: Channel Contrast b	A2 f7: Tong_BlurExtent
A1 f55: V DOF Ind	A2 f7: Tong_BlurExtent	A4 f9: Dark Channel	A1 f22: Size feat
A1 f1: mean intensity	A1 f43: P3 Relative Size	A6 f26: Sift FV Probability	A7 f13: Sharpness
A3 f2: Lightning	A8 f147: Eye Sensitivity 7	A4 f18: F2 Shadow area	A3 f10: Color Harmony 6
A2 f1: BBox_Edges	A9 f27: Itten S05	A2 f5: Blur	A1 f1: mean intensity
A2 f9: EdgeDist_Low	A2 f5: Blur	A6 f9: Channel Contrast b	A2 f5: Blur
A2 f5: Blur	A9 f34: Symmetry	A2 f9: EdgeDist_Low	A1 f22: Size feat
A8 f10: Dark Channel 5	A1 f7: ROT mean V	A2 f7: Tong_BlurExtent	A8 f1: Sal Region Area
A4 f9: Dark Channel	A1 f4: mean hue	A8 f92: White Balance 6	A6 f21: Noisiness
A1 f1: mean intensity	A8 f144: Eye Sensitivity 4	A4 f18: F2 Shadow area	A2 f7: Tong_BlurExtent
A1 f54: S DOF Ind	A2 f9: EdgeDist_Low	A2 f1: BBox_Edges	A4 f16: Hue Complexity
A1 f17: V WVT feat L2	A1 f43: P3 Relative Size	A9 f16: Itten H06	A4 f13: Spatial Complexity Fore
A6 f23: Dominant Color a	A1 f55: V DOF Ind	A1 f1: mean intensity	A4 f9: Dark Channel
A1 f7: ROT mean V	A1 f1: mean intensity	A1 f22: Size feat	A9 f16: Itten H06

TABLE 4: List of the top performing features for each of the four image categories of the HiddenBeauty dataset. Each row shows the algorithm number, based on the order presented in Section 3. and its description. Features on the bottom are the top-performing features without the Quality meta-features (NoDLFV).

All	Animals	Nature	People	Urban
0.665	0.678	0.508	0.592	0.617

TABLE 5: Comparison of the hand-crafted feature performance in predicting aesthetics score in terms of the correlation coefficients for the HiddenBeauty image categories.

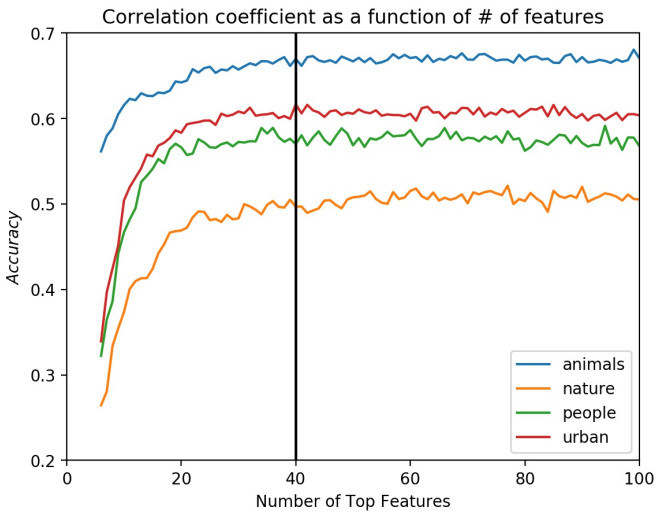


Fig. 4: Regression performance as the function of the number of top k features the different categories of the HB dataset. The vertical line at $k = 40$ indicated a point after which the regression performance remained approximately constant

features. Similarly to regression with all of the categories, the performance stays roughly the same, until the number of remaining featuring is ~ 40, after which we see a drop-off in the performance with decreasing number of features. It is interesting to observe the drop-offs for as the # of features is ≤ 20 - we can see that *people* and *urban*categories observe much smaller drop-off as compared to the *animals* and *people*. Table 4 shows the best performing features for the different categories with (top) or without (bottom) the quality meta-features. As can from the top part of the Table 4, in all of the categories the DL Probability (probability of being High-Quality image based on the CNN features) is the most informative feature for predicting continuous aesthetic score.

The bottom part of the Table 4 provides us with informations about the type of features that are important to predicting the scores for images in each category without higher-level meta features. It is observed that feature describing sharpness or technical quality of the image are important across all of the categories (e.g. features measure blur in [5], Wavelet-based features aiming to capture the Depth of Field (DOF) in [4]). Feature capturing different properties of color are observed to be among the most informative for all of the categories as well (e.g. Mean Hue and Hue Complexity). It is interesting to note that the some of the most important features in each category are very intuitive. For example, as we can see the single most important feature in the animals category is the first feature described in [5], which measures the normalized area of a bounding box enclosing 90% of the edge energy in the image. Such a feature is important, since it highlights images with a well-defined subjects (animals in the forefront of a blurred background). Often, many of the most appealing images of landscapes and flowers have vivid color - this notion is captured by the best performing feature for the Nature category is a bin in the Itten histogram [15], which measures the number of pixels in the image with high saturation. Similarly many of such images are very symmetric, as is captured by the Symmetry features, which measures as the absolute difference between the Histogram of Oriented Gradients (HOG) descriptors of the left and right halves of the image [10]. Lastly for the Human category, some of the most important features measure the amount of shadow area on the faces (as picked out by a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2845100, IEEE Transactions on Image Processing

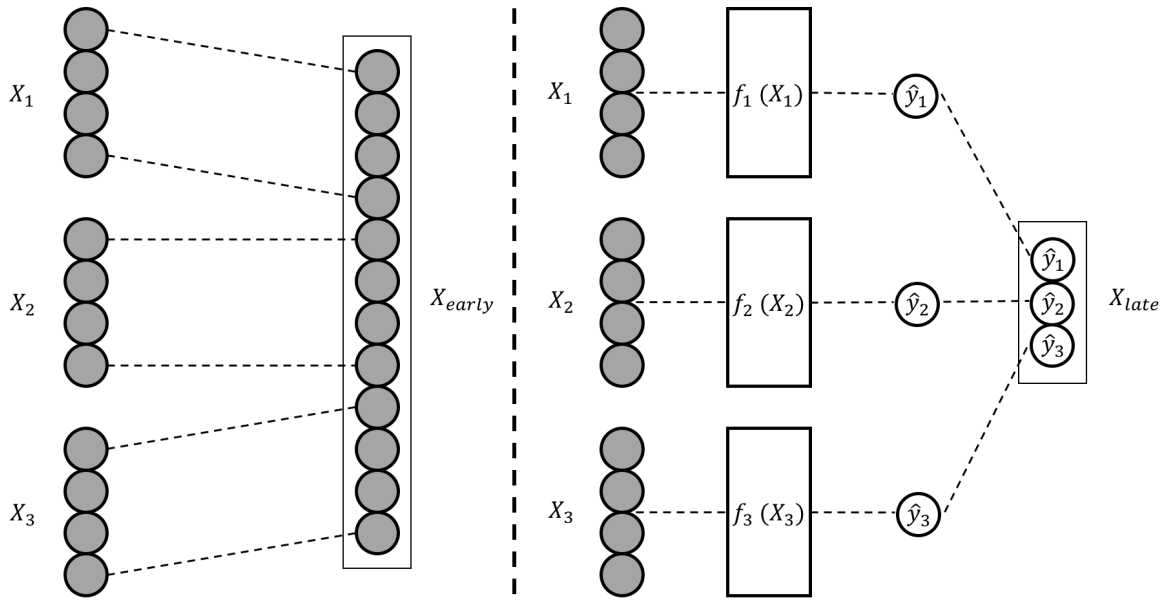


Fig. 5: Structure of the general pipeline, where we concatenate the HC features with CNN activations.

pre-trained face detector) and white balance in the image corroborating the notion that lighting of the faces affects our perception of the image [9].

4 Combining the CNN and HC features

In this section, we investigate the possibility of improving the performance of deep learning models by fusing hand-crafted features with CNN activations from the penultimate layers of the networks and use them to predict both the mean aesthetics score on the HB dataset, and predicting High / Low Quality images in the AVA dataset. In combining the HC features, we consider the top performing hand-crafted features after feature elimination based on the assumption that they will provide the most discriminatory power (we examine both inclusion and exclusion of the DL Probability and Fisher Vector features in the respective combinations).

4.1 Choosing baseline CNN features

In this section, we describe the comparison of a sample of popular baseline CNN architectures, of which we choose two to be combined with HC features. As can be seen in the recent review by Deng et al. [14], many of the baseline models and proposed architectures are based on the popular AlexNet [13], which first achieved state-of-the-art results on the ImageNet competition. We perform a baseline comparison of the following four models: VGG16 [32], VGG19 [32], ResNet50 [36], and Inception [37]. We provide this comparison in order to choose a strong model to compare against the HC features in the following section and to avoid biasing our results by comparing them to weak baseline models.

Experimental Setup

In order to evaluate the different baseline CNN models, we use the CUHKPQ dataset, where we predict the High/Low quality of the images. In estimating the performance of the algorithms, we utilize 80-20 training-testing splits for classification. In order to better estimate the score, we take the mean of 20 trials, where we randomly perform a split and calculate the respective metrics. For classification, we report the overall accuracy, defined as

$$\text{Overall Accuracy} = \frac{TP + TN}{P + N}, \tag{5}$$

where TP is the number of true positive examples, TN is the number true negative examples and $P + N$ is the total number of images.

VGG16	VGG19	ResNet50	InceptionV3
0.918	0.920	0.936	0.894

TABLE 6: Classification performance of the CUHKPQ dataset on the baseline CNN features for CNN models pre-trained on the ImageNet dataset.

All of the models evaluated were top performers in the ImageNet competition, thus, they provided good baseline results both in terms of classification accuracy and regression. Table 6 shows that, even baseline features from the pen-ultimate layers of all of the models do a reasonably good job in predicting the quality of images in the CUHKPQ dataset as compared to summarized results in [14], with ResNet50 model achieving the best performance and InceptionV3 achieving the worst and therefore will not be considered further.

4.2 Improving CNN performance with HC features.

Section 3.3 considered a way of combining HC and CNN features, by using CNN activations to construct a model to predict the meta-feature indicating the probability of the image being High-Quality based on the trainings set of images in the CUHKPQ dataset. This meta-feature is then considered as one of the HC features. In this section, we

Feature	Description
A2: BBox Edges	measures the bounding box enclosing the 90% of the edge energy
A2: Blur	measure of blur in an image as defined in [5]
A2: Tong_Per	measure of blur in an image as defined in [38]
A6: Channel Contrast L	computes the width of the middle 90% mass of the L channel in the Lab space
A6: ColorComp V2	predicts the probability that the picture is of High Quality
A6: Sharpness	FFT based measure of sharpness of the given image
A8: Sal Region Area	total area of a binary-thresholded saliency map
A8: ROT	Distances to the four stress points in the image from the subject of the image
A8: Colorfulness	Color Hue count in a given image region
A8: Color Harmony	Color Harmony descriptor as defined in Nishiyama et al. [11]
A9: Contrast	measures the normalized difference between the minimiun and maximum luminance
A9: Arousal	Indicator of Emotion that combines average Saturation and Brightness as defined in [15]
A9: Itten	HSV color histogram described in [10]
A9: Symmetry	measures the similariy between the left and right halves of an image [10]
A9: Contrast	Haralick's Contrast features based on the Gray-Level Co-occurence Matrix [39]
A9: Entropy	Haralick's Entropy features based on the Gray-Level Co-occurence Matrix [39]

TABLE 7: List and description of the the top performing HC features / sets of features that improve the performance of CNN models.

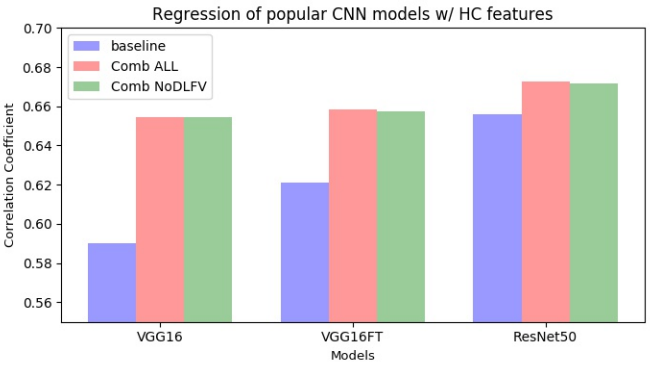


Fig. 6: Regression performance on the HiddenBeauty score for various CNN models and their combination with HC features.

Previous work	Overall Accuracy
AVA handcrafted features (2012) [22]	68.00
Kao et al. (2016) [40]	74.51
RAPID - improved version (2015) [17]	75.42
DMA net (2015) [41]	75.41
Kao et al. (2016) [42]	76.15
Wang et al. (2016) [43]	76.94
Kong et al. (2016) [20]	77.33
Mai et al. (2016) [44]	77.40
BDN (2016) [18]	78.08
ILGNet (2017) [45]	79.95
VGG16	79.41
VGG16 Early Fusion	80.83
VGG16 Late Fusion	81.65
ResNet50	81.27
ResNet50 Early Fusion	81.79
ResNet50 Late Fusion	81.95

TABLE 8: Classification performance of different models on the AVA dataset.

	VGG	**ResNet**
$\mathcal{H}_1 : \pi_A \neq \pi_B$	1.176e−8	0.01482
$\mathcal{H}_1 : \pi_A < \pi_B$	5.5881e−9	0.00741

TABLE 9: The following table shows the p-values for the one-sided and two-sided McNemar Test at the significance value of $\alpha = 0.05$

explore two ways of fusing hand-crafted features $\mathbf{X}_{HC}$ and learned CNN activations $\mathbf{X}_{CNN}$ from the penultimate layer of the network: early (classification / regression) and late fusion (classification). In early fusion, the HC features are concatenated with the CNN features into a single features representation

$$\mathbf{X}_{early} = [\mathbf{X}_{HC}; \mathbf{X}_{CNN}]$$

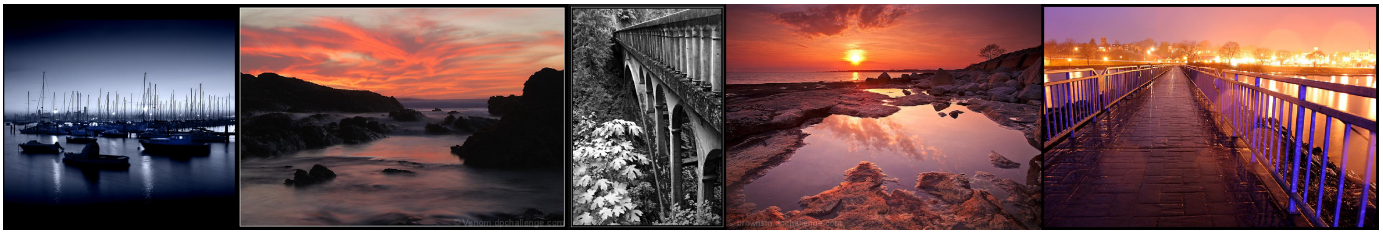
, which are then used to learn a function $f : \mathbf{X}_{early} \rightarrow \mathcal{Y}$. Alternatively, we explore a late decision-level fusion by model stacking, where we learn two levels of models (this necessitates splitting the training set feature for the particular dataset into two parts). In the first level, we learn separate models based f_{HC} and f_{CNN} based on HC and CNN feature representations respectively (using the first part of each $\mathbf{X}_{HC}$ and $\mathbf{X}_{CNN}$). Then we use the second model

$$f_{stack} : [f_{HC}; f_{CNN}] \rightarrow \mathcal{Y}$$

to learn a function to combine the decisions of the first-level models. As we will show, both early and late fusion approaches on average improve the performance in prediction of both the mean aesthetics score and binary classification (see Figure 5).

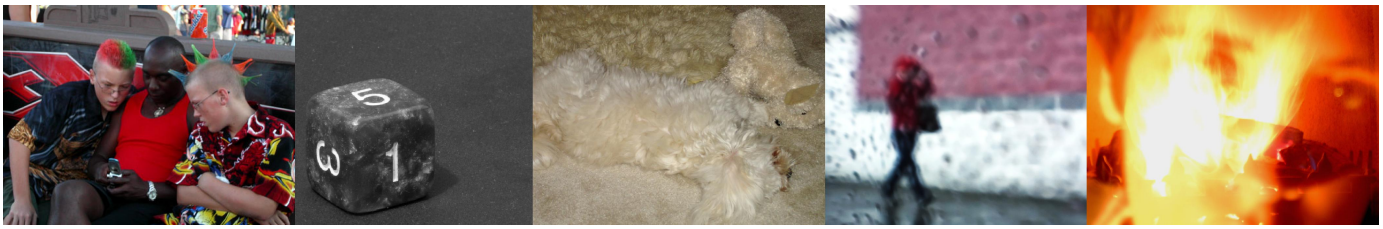
Figure 6 shows the summary of results for predicting mean aesthetics score, where the "baseline" R^2 comes only from the CNN features, whereas "Comb ALL/NoDLFV" combine the HC and CNN features[2]. As we can see from our results, simply concatenating HC features and CNN gives a more significant improvement as compared to using CNN features as a meta-feature. Although fine-tuning VGG16 to predict binary scores on the CUHKPQ dataset does provide better baseline features, the baseline features are combined with HC features, we see a very negligible difference between the two models. Additionally, including the meta-features predicting the quality of the images based on CNN and FV provides little improvement. Figure 6 shows that

2. The William's test for the difference between correlated correlations return a value of $t = -5.41$ and $p = 6.2\text{e}{-8} < \alpha = 0.05$ for the regression performance before and after adding HC features to ResNet CNN features, suggesting this improvement is statistically significant.



(a) Top correctly classified images of High Quality

(b) Top correctly classified images of Low Quality



(c) Incorrectly classified images of High Quality

(d) Incorrectly classified images of Low Quality

(e) Images of High Quality that were correctly classified by concatenating HC features.

(f) Images of Low Quality that were correctly classified by concatenating HC features.

Fig. 7: Sample images from the AVA dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2845100, IEEE Transactions on Image Processing

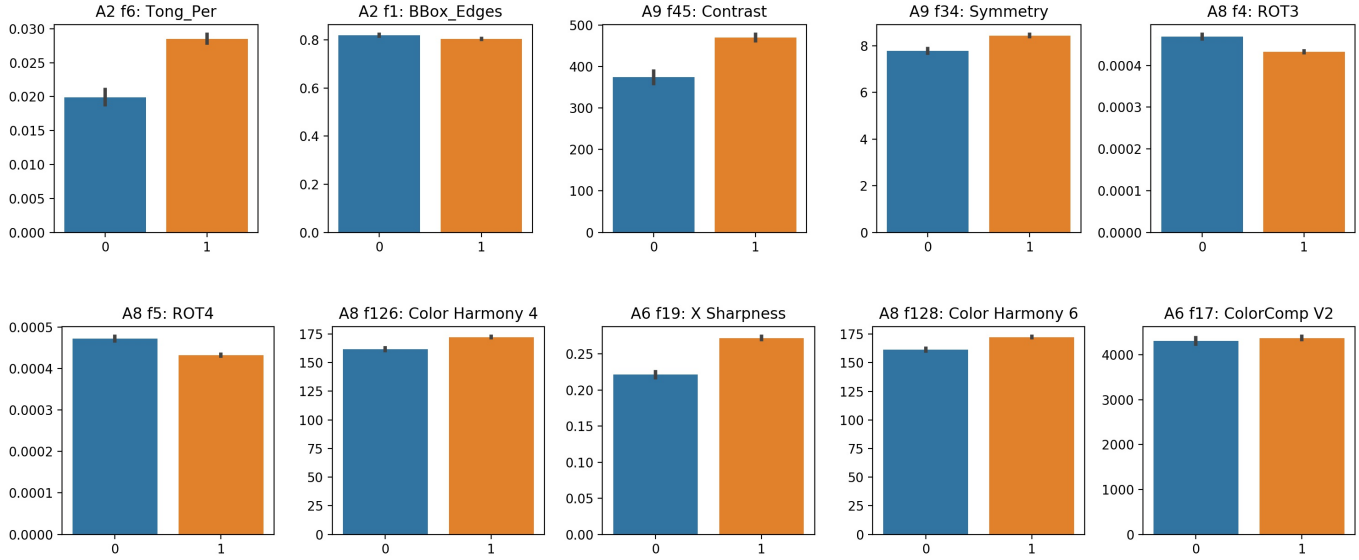


Fig. 8: Plot of the distribution of the top performing hand-crafted features across the high and low quality classes.

Model	ResNet	VGG16
1	A2 f6: Tong_Per	A2 f6: Tong_Per
2	A2 f1: BBox_Edges	A6 f19: X Sharpness
3	A8 f126: Color Harmony 4	A2 f1: BBox_Edges
4	A8 f128: Color Harmony 6	A8 f126: Color Harmony 4
5	A9 f34: Symmetry	A8 f128: Color Harmony 6
6	A6 f19: X Sharpness	A6 f20: Y Sharpness
7	A9 f1: Contrast	A8 f140: Color Harmony 18
8	A9 f45: Contrast	A9 f34: Symmetry
9	A8 f5: ROT4	A9 f45: Contrast
10	A8 f4: ROT3	A8 f4: ROT3
11	A6 f17: ColorComp V2	A9 f1: Contrast
12	A8 f130: Color Harmony 8	A8 f132: Color Harmony 10
13	A8 f132: Color Harmony 10	A8 f114: Colorfulness 1
14	A9 f9: Arousal	A8 f5: ROT4
15	A9 f27: Itten S05	A9 f16: Itten H06

TABLE 10: Top 15 performing Hand-Crafted features for the models combining HC and pre-trained CNN features.

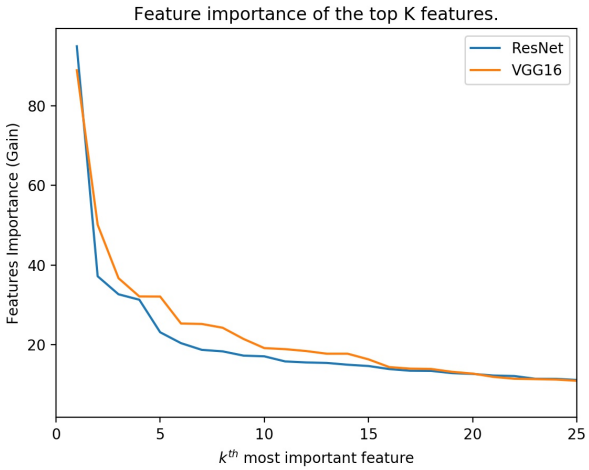


Fig. 9: Feature Importance (Gain) of the k^{th} feature.

ResNet50 achieves better results than VGG16 as well as achieving a smaller improvement when concatenating it with HC features.

In order to quantify the improvement HC features can provide in a real world scenario, we train a binary classifier based on early and late fusion approaches described earlier to predict the High/Low Quality of Images on the AVA dataset. Similar to strategy used in previous papers [17], [22], [45], the AVA dataset is split into high / low quality images by assigning those images with a score ≥ 5 to the high quality class with roughly 235,000 images being used for training and 20,000 images for testing[3].

Table 8 lists the performance of the various algorithms. Although both of the baseline networks achieve very good performance in predicting the binary aesthetics, an improvement of up to 2.2 % can be achieved by fusing the network features with HC features, with decision-level fusion achieving a bigger improvement as compared to feature-level fusion. The Table 9 shows us the p-values for comparing the classifiers A (baseline) and B (fusion) tested at the significance level of $\alpha = 0$. As can be seen, in each case we have that $p-$values $< \alpha$, indicating we should reject the null hypothesis $\mathcal{H}_0$ at the 5% significance level, suggesting that the fusion of CNN and HC features does indeed improve the performance of the models.

Figure 7 shows sample images that were correctly classified by the models with combined features (Figure 7 (a) and (b)), misclassified (Figure 7 (c) and (d)), and images classified by the models with the combined features but misclassified by the models based only on the pre-trained features (Figure 7 (e) and (f)).

Since each baseline achieves a significant improvement after feature fusion, we can examine the top-performing HC features for the different models and understand, for example, the type of high level knowledge that is approximated by the HC features and missing from CNN, as well

3. The particular training/testing splits are the same as ones used in [45] and can be found at: https://github.com/BestiVictory/ILGnet

as the differences between the various CNN models. Table 10 shows the fifteen most important hand-crafted features (note that among the top 100 features as ranked by the gain-based ranking criterion, 30 and 49 features of them are hand-crafted for the RESNET and VGG16 model respectively) and Figure 9 shows the plot of the feature importance as quantified by the gain in descending order for the top features. It can be seen that both of the models use very similar features in improving aesthetics classification, which include features that relate to photographic rules used by professional photographers, e.g. Symmetry features of A9 or features that measure sharpness / blur of photographs. Despite both of the feature sets using features that capture similar information, HC features have a larger impact in improving the performance of the model when concatenated with VGG features as opposed to ResNet features (HC features improve VGG model by 2.2% as opposed to 0.7 % with ResNet).

To gain a better understanding of the image that were correctly classified with HC features, in Figure 8 we plot and examine the distribution of the values across the high and low quality classes of the top performing features as judged by the model. The best performing HC feature for both of the models is the Wavelet-based *Per* feature defined in [46] measuring blur, where an image is said to be un-blurred if *Per* is greater than some threshold. We can see from the first plot that images of high quality indeed have a higher value *Per* and thus are less "blurry". Similarly, *BBox_edges* feature estimates the size of the bounding box enclosing 90 % of edge energy in the image [5]. Intuitively, if the image has a defined subject, most of the edge energy should be concentrated within a smaller box, which is indeed true. The ROT3 feature estimates the normalized distance from the center of mass of a saliency map to the anchor points of an image (see Figure 1). As we can see, the images with higher quality tend to have lower distance to one of the anchor points, suggesting a better adherence to the rule of thirds in photography composition.

5 Conclusion

In this paper, we studied and compared a selection of algorithms that use hand-crafted features designed to assess image aesthetics. We show that even early algorithms can provide adequate results in their efficacy of predicting image aesthetics as compared to more recent methods based on hand-crafted features. We can achieve an additional improvement in aesthetic prediction accuracy by combining all of the features together and attain a performance close to that of a model trained on learned CNN features. By performing feature elimination, a good performance for classification / regression can be achieved for a specific combination of just 25 and 40 features respectively out of total more than 300 features. Furthermore, we can see that even if we remove a large portion of features (in our case these were features from Algorithm 8), we achieve a very similar performance with features from different algorithms that captured very similar information (e.g. Color Harmony from A4 vs A8). By analyzing the combination of all features on the different categories, we find that the most important features of each category are intuitively important for each of the categories. Furthermore, when fusing HC features with pre-trained deep learning features, we can achieve a significant improvement in predicting both a continuous aesthetic metric, and predicting a binary high / low quality score with improvement up to 2.2 % in classification accuracy.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

Appendix

The McNemar's Test [4] is a hypothesis test for comparing populations proportions, considering the fact that the data come from two dependent matched-pair samples [47] (i.e. the predictions of the classifiers are trained and tested on the same training / testing splits). Assuming a learners A and B and their corresponding decisions functions $f_A(x)$ and $f_B(x)$ were trained on the same training set, let $\{\hat{y}_i^A\}$ $\{\hat{y}_i^A\}$ be the predictions for the test set obtained from the learners A and B respectively. Then the two-sided test for comparing the accuracies of the models is:

$$\mathcal{H}_0 : \pi_A = \pi_B$$
$$\mathcal{H}_1 : \pi_A \neq \pi_B$$

where pi_i represents the misclassification rates of the two models (in our case, the model A corresponds to the model trained will solely deep features B is the model that combines the CNN features with HC features). Alternatively, we can test the the the alternative hypothesis $\mathcal{H}_1 : \pi_A < \pi_B$.

References

[1] S. Heyman, "Photos, photos everywhere," Jul 2015. [Online]. Available: https://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html
[2] T. Ang, *Digital Photographer's Handbook*, 5th ed. DK ADULT, 2012.
[3] J. Zuckerman. (2017) Jim zuckerman on composition: Symmetry. [Online]. Available: https://www.photovideoedu.com/Learn/Articles/jim-zuckerman-on-composition-symmetry.aspx
[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, *Studying Aesthetics in Photographic Images Using a Computational Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 288–301.
[5] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 419–426.
[6] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proceedings of the 10th European Conference on Computer Vision: Part III*, ser. ECCV '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 386–399.
[7] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, Nov 2009, pp. 997–1000.
[8] S. Dhar, V. Ordonez, and T. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 1657–1664.
[9] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *Multimedia, IEEE Transactions on*, vol. 15, no. 8, pp. 1930–1943, Dec 2013. [Online]. Available: http://dx.doi.org/10.1109/TMM.2013.2269899

4. See MATLAB function *testcholdout* for an implementation of the algorithm.

[10] R. Schifanella, M. Redi, and L. M. Aiello, "An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures," in *ICWSM'15: Proceedings of the 9th AAAI International Conference on Weblogs and Social Media*. AAAI, 2015.

[11] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 33–40. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995539

[12] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1784–1791. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2011.6126444

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[14] Y. Deng, C. C. Loy, and X. Tan, "Image aesthetic assessment: An experimental survey," *CoRR*, vol. abs/1610.00838, 2016. [Online]. Available: http://arxiv.org/abs/1610.00838

[15] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 83–92. [Online]. Available: http://doi.acm.org/10.1145/1873951.1873965

[16] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, "Towards a comprehensive computational model for aesthetic assessment of videos," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 361–364. [Online]. Available: http://doi.acm.org/10.1145/2502081.2508119

[17] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 457–466. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654927

[18] Z. Wang, F. Dolcos, D. Beck, S. Chang, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," *ArXiv e-prints*, Jan. 2016.

[19] L. Guo and F. Li, "Image aesthetic evaluation using paralleled deep convolution neural network," *CoRR*, vol. abs/1505.05225, 2015. [Online]. Available: http://arxiv.org/abs/1505.05225

[20] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision (ECCV)*, 2016.

[21] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan-Kaufmann, 1994, pp. 737–744. [Online]. Available: http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf

[22] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," pp. 2408–2415, June 2012.

[23] W. Jiang, A. C. Loui, and C. D. Cerosaletti, "Automatic aesthetic value assessment in photographic images," in *2010 IEEE International Conference on Multimedia and Expo*, July 2010, pp. 920–925.

[24] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li, "The new data and new challenges in multimedia research," *CoRR*, vol. abs/1503.01817, 2015. [Online]. Available: http://arxiv.org/abs/1503.01817

[25] F. Liu and S. Osindero, "A complete framework for aesthetic inference in images," *arXiv*, 2015. [Online]. Available: http://stanford.edu/~liuf/papers/cvpr2015_sub1.pdf

[26] F. Simond, N. Arvanitopoulos, and S. Ssstrunk, "Image aesthetics depends on context," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 3788–3792.

[27] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[28] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[30] DMLC, *Understand your dataset with XGBoost*, 2016 (accessed September 20, 2016). [Online]. Available: http://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html

[31] R. Kohavi, "Wrappers for performance enhancement and oblivious decision graphs," Ph.D. dissertation, Stanford, CA, USA, 1996, uMI Order No. GAX96-11989.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[33] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan 2002. [Online]. Available: https://doi.org/10.1023/A:1012487302797

[34] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273 – 324, 1997, relevance. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000437029700043X

[35] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 431–439. [Online]. Available: http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[38] H. Tong, M. Li, H. Zhang, and C. Zhang, "Blur detection for digital images using wavelet transform," in *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, vol. 1, June 2004, pp. 17–20 Vol.1.

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

[40] Y. Kao, K. Huang, and S. Maybank, "Hierarchical aesthetic quality assessment using deep convolutional neural networks," *Image Commun.*, vol. 47, no. C, Sep. 2016.

[41] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 990–998.

[42] Y. Kao, R. He, and K. Huang, "Visual aesthetic quality assessment with multi-task deep learning," *CoRR*, vol. abs/1604.04970, 2016.

[43] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Image Commun.*, vol. 47, no. C, pp. 511–518, Sep. 2016. [Online]. Available: https://doi.org/10.1016/j.image.2016.05.009

[44] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 497–506.

[45] X. Jin, J. Chi, S. Peng, Y. Tian, and C. Y. andXiaodong Li, "Deep image aesthetics classification using inception modules and fine-tuning connected layer," in *8th International Conference on Wireless Communications & Signal Processing, WCSP 2016, Yangzhou, China, October 13-15, 2016*, 2016, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/WCSP.2016.7752571

[46] H. Tong, M. Li, H. Zhang, and C. Zhang, "Blur detection for digital images using wavelet transform," in *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, vol. 1, June 2004, pp. 17–20 Vol.1.

[47] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998. [Online]. Available: https://doi.org/10.1162/089976698300017197

Michal Kucer Michal Kucer received his BS in Microelectronic Engineer-

ing and Applied Mathematics from the Rochester Institute of Technology (2014) and is currently pursuing a PhD in Imaging Science at RIT. His work focuses on the development of methods for predicting the aesthetic value of images. His broader interests include Computer Vision, Remote Sensing and Machine Learning.

Alexander C. Loui Alexander Loui received his Ph.D. (1990) in Electrical Engineering from the University of Toronto, Canada. He is currently a Technical Lead and Senior Principal Scientist at Kodak Alaris in Rochester, NY. He is also an Adjunct Professor of ECE Department at Ryerson University and University of Toronto. He has been directing research on computer vision, video summarization, machine learning, image aesthetics, image event analysis, and multimedia applications. He is a Fellow of IEEE and SPIE.

David W. Messinger David W. Messinger received a Bachelors degree in Physics from Clarkson University and a Ph.D. in Physics from Rensselaer Polytechnic Institute. He is currently a Professor, the Xerox Chair in Imaging Science, and Director of the Chester F. Carlson Center for Imaging Science at the Rochester Institute of Technology. His personal research focuses on projects related to remotely sensed spectral image exploitation using physics-based approaches and advanced mathematical techniques.