A Holistic Approach to Aesthetic Enhancement of Photographs

SUBHABRATA BHATTACHARYA, University of Central Florida
RAHUL SUKTHANKAR, Google Research
MUBARAK SHAH, University of Central Florida

This article presents an interactive application that enables users to improve the visual aesthetics of their digital photographs using several novel spatial recompositing techniques. This work differs from earlier efforts in two important aspects: (1) it focuses on both photo quality assessment and improvement in an integrated fashion, (2) it enables the user to make informed decisions about improving the composition of a photograph. The tool facilitates interactive selection of one or more than one foreground objects present in a given composition, and the system presents recommendations for where it can be relocated in a manner that optimizes a learned aesthetic metric while obeying semantic constraints. For photographic compositions that lack a distinct foreground object, the tool provides the user with crop or expansion recommendations that improve the aesthetic appeal by equalizing the distribution of visual weights between semantically different regions. The recomposition techniques presented in the article emphasize learning support vector regression models that capture visual aesthetics from user data and seek to optimize this metric iteratively to increase the image appeal. The tool demonstrates promising aesthetic assessment and enhancement results on variety of images and provides insightful directions towards future research.

Categories and Subject Descriptors: H.4.m [**Information Systems Applications**] Miscellaneous

General Terms: Algorithms, Human Factors

Additional Key Words and Phrases: Interactive photo tools, spatial recomposition, quality enhancement

ACM Reference Format:

Bhattacharya, S., Sukthankar, R., and Shah, M. 2011. A holistic approach to aesthetic enhancement of photographs. ACM Trans. Multimedia Comput. Commun. Appl. 7S, 1, Article 21 (October 2011), 21 pages.
DOI = 10.1145/2037676.2037678 http://doi.acm.org/10.1145/2037676.2037678

1. INTRODUCTION

The last decade has seen a deluge of image sharing Web sites owing to the increasing popularity and affordability of consumer grade digital cameras and ease of accessibility of digital images. Needless to say, an ordinary camera user is no longer selective while taking photographs. From a photo sharing users' perspective, this has introduced two new problems: the first is the ability to select the best-looking ones from a large pool of photographs captured during certain occasion. The next is the flexibility to edit a photograph with minimal photographic compositional knowledge so that the result

Author's address: S. Bhattacharya; email: subh@cs.ucf.edu.

DOI 10.1145/2037676.2037678 http://doi.acm.org/10.1145/2037676.2037678

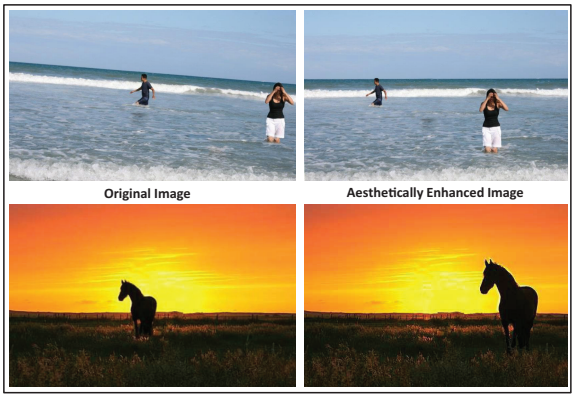


Fig. 1. Photo-quality enhancement. The images on the left are input to our composition enhancement tool, while their enhanced counterparts are shown on the right.

looks reasonably better than the original ones. These two key issues have paved the path for interesting research in photographic quality assessment and enhancement.

The concept of a "high quality" image as perceived by a viewer is often abstract, even for professional photographers, which is why assessing the aesthetic quality of photographs is challenging. However, photographs taken by experienced photographers, similar to paintings by renowned artists, adhere to several rules of composition that make them more visually appealing than those taken by amateurs. Studies have revealed that such photographic compositions trigger several psycho-visual stimuli in the human observer due to which the photograph is perceived to be of good quality. As described in the photography literature [Jonas 1976], these include the *Rule of Thirds* and *Visual Weight Balance*. Elementary photography lessons emphasize that adhering to these two rules alone could significantly improve the aesthetic quality of most amateur photographs. According to the *Rule of Thirds*, the photographer or artist should place the primary subject of the composition near a location that is a strong focal point. Similarly, in order to conform to the rule of *Equalized Visual Weights*, in a well-composed image the visual weights of different regions should be in proportion to the *Golden Ratio*. A simple heuristic such as keeping the horizon as parallel as possible to the horizontal axis of the photographic frame also generally increases the aesthetic appeal of an image (Figure 1). Figure 1 shows two results of our photocomposition enhancement algorithms in two natural photos. In order to make this article self-contained, we discuss these compositional rules with practical examples in detail.

We organize this article as follows. In the next section, we discuss the body of literature close to the problem we are addressing in this article. This section is concluded by an overview of our assessment and enhancement engine, which are the two major components of our photographic composition enhancement tool. In the following section, we present the technical details involved in learning and evaluating the aesthetic quality of a photograph, along with results demonstrating its agreement with human ratings. In Section 5, we describe our approach to enhance the aesthetic appeal of photographs through the proposed recomposition framework and show examples from our dataset that highlight specific aspects of the process. This is followed by experimental results on assessment and recomposition. Finally, we conclude this article by discussing several possible applications.

2. RELATED WORK

Research in evaluating photographic quality dates back to the work of Venkata et al. [2000], who use a reference image with its noise degraded counterpart to assess its quality. More recently, Ke

			Aspects		
Method	Assessment	Metric	Enhancement	Scene Semantics	Dataset Source
[Boutell and Luo 2004]	Addresses	Low-level	-	-	Personal
[Ke et al. 2006]	Addresses	Low/Mid-level	-	-	DPchallenge.com
[Datta et al. 2006]	Addresses	Low/Mid-level	-	-	Photo.net
[Avidan and Shamir 2007]	-	Low-level	Addresses	-	Personal
[Datta et al. 2007]	Addresses	Low/Mid-level	-	-	Photo.net
[Luo and Tang 2008]	Addresses	High-level	-	-	DPchallenge.com
[Cho et al. 2008]	-	Low-level	Addresses	-	Personal
[Nishiyama et al. 2009]	-	Low-level	Addresses	-	Personal
[You et al. 2009]	Addresses	Low/Mid-level	-	-	Photo.net
Proposed	Addresses	High-level	Addresses	Addresses	Flickr.com

Fig. 2. Comparison of the proposed photo-composition assessment/enhancement approach with previously published methods. Refer to the text for detailed description of the various aspects.

et al. [2006] construct high-level features for photo quality assessment extracted from low level cues like noise, blur, color, brightness, contrast and spatial distribution of edges. In addition to some of these low-level cues, [Datta et al. 2006, 2007] investigate the impact of features such as familiarity measures, wavelet responses on textures, aspect ratio, and region composition on the aesthetic appeal of natural images. Boutell and Luo [2004] explore a variety of metrics including ISO speed rating, F-number and shutter speed, extracted directly from camera metadata, to determine their impact on photographic quality. These methods, as observed by Luo and Tang [2008] and Sun et al. [2009], capture only fine-grained details about the photograph that are mainly introduced due to sensors used during the image formation process. Thus, in order to understand the nuances of spatial composition in photographic frames, Luo and Tang [2008] additionally introduced a parameter that considered adherence to geometric composition rules for photo and video quality evaluation.

While Luo and Tang [2008] demonstrated some level of success in evaluating photo quality in natural photographs, that approach relies heavily on a blur detection technique to identify the foreground object's boundary within the frame. This technique works well only with photographs captured using professional SLR cameras that have mechanisms to induce depth-of-field effects and precludes its use with photographs taken using popular point-and-shoot cameras. The interested reader is requested to refer to Figure 2 for a summary of existing approaches in this field in terms of different characteristic properties.

We argue that true aesthetic assessment should not be constrained by equipment capability as photographs captured using professional cameras are rarer in number and often restricted by terms of use. Furthermore, photographs captured using professional equipment are more likely to follow composition guidelines since they are generally taken by experienced photographers. Our approach can be characterized as a method to improve photographs, such as those frequently found on the Internet, that were taken by amateurs using consumer digital cameras.

3. APPROACH OVERVIEW

We formulate photo quality evaluation as a machine learning problem in which we map the characteristics of a human-rated photograph in terms of its underlying adherence to the rules of composition. A part of our method can be compared with the approach suggested in Sun et al. [2009], You et al. [2009], and Mansoor et al. [2009], wherein the authors apply a saliency map to estimate visual attention distribution in photographs. We complement the saliency information extracted from an image using a high-level semantic segmentation technique that infers the geometric context [Hoiem et al. 2007] of a scene. With the help of the above methods, we extract aesthetic features that could be used to measure

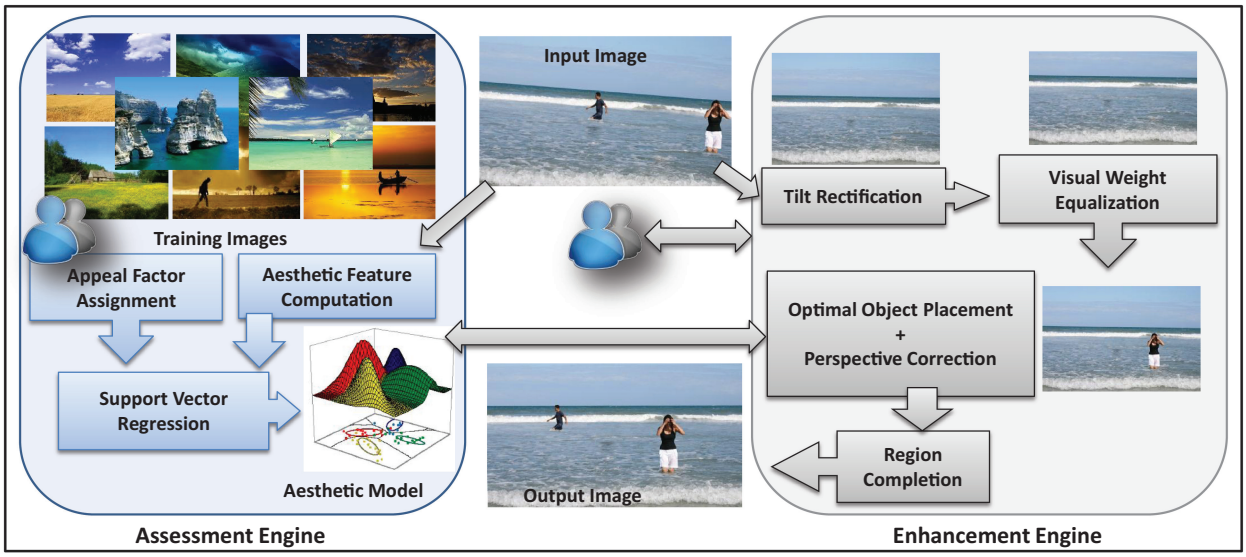


Fig. 3. Schematic diagram of our photo enhancement application: The tool consists of two components: the Assessment Engine (AE) and the Enhancement Engine (EE). The primary objective of the AE is to learn a mapping between measured features in images with their respective appeal factors obtained from an extensive user study. EE on the other hand, is responsible for generating recompositions that optimize the predicted appeal factor of an input image while obeying semantic constraints.

the deviation of a typical composition from ideal photographic rules of composition. These aesthetic features are subsequently used as input to two independent Support Vector Regressors in order to learn the visual aesthetic model. This learned model is then integrated into our photo-composition enhancement framework. To this end, we make the following contributions in this article: (1) perform an empirical study on visual aesthetics using real human subjects on real-world images; (2) find a smooth mapping between user input visual attractiveness and high-level aesthetic features; (3) apply semantic scene constraints while recompositing a photograph; (4) introduce an interactive tool that helps users to recompose photographs with some informed aesthetic feedback; and (5) bring photographic quality assessment and enhancement under a single unifying framework. An overview of our approach is shown in Figure 3.

We primarily focus on outdoor photographic compositions with one or more foreground subjects or compositions with no dominant foreground subjects. For the former, we constrain our algorithm to relocate the objects to a more aesthetically pleasing location while respecting the scene semantics (e.g., a tree attached to the ground must remain in contact with the ground) and rescaling it as necessary to maintain the scene's perspective. This is a significant improvement over a foreground object-centric image-editing technique [Cho et al. 2008], wherein the authors propose a method to reconstruct an image from low-resolution patches subject to user-defined constraints. We also show that how our technique can be extended to handle multiple foreground objects. In the case of photographs that lack a dominant subject, such as land/seascapes, we crop or expand the photograph so that an aesthetically pleasing balance between sky and land/sea is achieved.

We demonstrate that these composition techniques can be used in combination with each other where horizon information is available. During the enhancement process, we automatically rectify tilt artifacts that may be present in the image. When the spatial alterations create holes where

the original photograph lacks information, we apply inpainting to preserve the photo-realism of the original while minimizing artifacts. In this context, our work is partially motivated by the work of Nishiyama et al. [2009], which introduces a method for automatically cropping a photo using a quality classifier built from user responses; their method implicitly assumes that in a given image, the background region is blurred to emphasize the subject region. Since we rely on a segmentation algorithm that provides us with semantic information of the scene, we can address a broader spectrum of photographs, relaxing this assumption. Our approach is also philosophically similar to Leyvand et al. [2008]'s work on beautification of human facial images, which quantifies the attractiveness of a human face from the spatial location of features such as eyes, lips, and nose, and alters the photograph so as to realign these features to more desirable positions. With this, we proceed towards the next section of the paper that discusses the two important components of our photo-composition enhancement workflow.

4. ASSESSMENT ENGINE

The assessment engine is responsible for providing the notion of quality or aesthetic metrics to the photo composition framework. For the simplest case of a single subject composition photograph, we use visual saliency [Walther and Koch 2006]-based techniques [Sun et al. 2009; You et al. 2009] to obtain a reasonable estimate of the spatial location of the dominant foreground regions in photographs. While this approach addresses our need for identifying the spatial location of the object in a photo-frame, it does not provide any scene semantics that we require to (1) assess the aesthetic appeal based on visual weight, or (2) recompose the given image while maintaining the scene integrity. We tackle this problem using a supervised learning-based scene classification method proposed by Hoiem et al. [2007]. This technique generates a confidence map of semantic labels that we can employ to identify likely regions of *sky* and *support* (a generic term for nonforeground regions that do not belong to sky) in an image. Since the images in our dataset are primarily single-subject compositions, the complementary regions in the image that belong to neither the sky nor support, correspond to the foreground (by rule of elimination). We use morphological processing tools to disregard small disconnected regions in order to obtain a reasonable mask for the foreground. Our tool allows users to interactively refine the foreground segmentation and horizon estimation, which is crucial to achieving aesthetically pleasing yet semantically correct results.

4.1 Dataset

In order to contribute to the research community, we make an attempt to build the first dataset of this kind which is reusable, expandable and publicly available. Our dataset[1] consists of 650 digital photographs, all downloaded from free image sharing portals, such as Flickr. Out of these, 384 images conform to the category of single-subject compositions, 18 multiple subject compositions, while the rest are of landscapes or seascapes that do not have any distinct foregrounds. Some of these images have artifacts that are commonly seen in photographs captured by novice photographers, such as improperly framed subjects or a tilted horizon. The compositions also exhibit a large variance in terms of spatial resolution. For computational efficiency, we have rescaled most of our images approximately equal to 640×480 or 480×640 depending on the nature of composition(landscape or portrait), keeping the aspect ratio constant. Figure 4 presents a subset of images that we have used in this paper. A Ground Truth aesthetic appeal factor (discussed in Sec. 4.2), associated with each image is used to evaluate the performance of our quality assessment algorithm and is used later to perform the recomposition.

[1]http://www.ucf.edu/∼subh/photoquality.

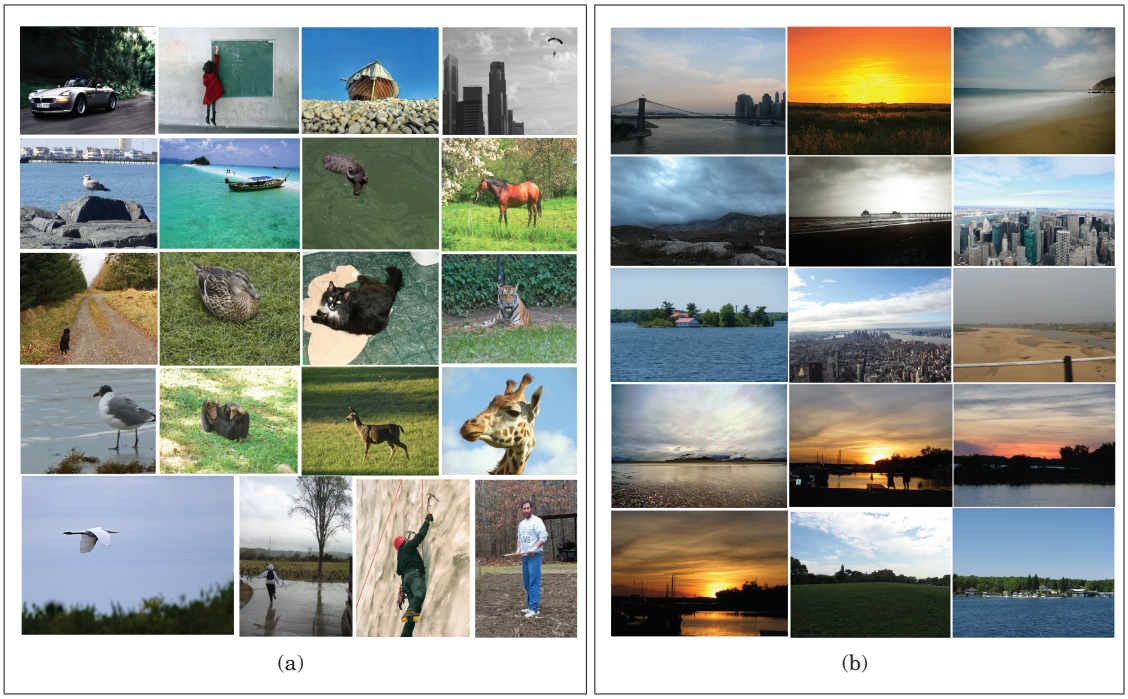


Fig. 4. A small sample of images from our dataset: (a) Compositions with a dominant foreground object, (b) Landscapes or seascapes without a foreground subject. [Acknowledgement: This dataset is based on copyright-free images shared from Flickr.]

4.2 User Survey

We conducted a thorough study of human aesthetics through a survey where 15 independent participants were asked to assign integer ranks to the photographs in our dataset from 1 to 5, with 5 being assigned to the most appealing. Further, while ranking, users were specifically instructed to eliminate bias from their ratings that might have emerged due to individual subject matter contained in a photograph, for instance, whether a user prefers mountains to sea or birds to animals. Each user was asked to rank no more than 30 images in a particular sitting in order to avoid undesirable variances in the ranking system due to fatigue or boredom. This process was further repeated 5 times to eliminate rankings from inconsistent users. After discarding the scores assigned by inconsistent users, we observed that the distributions were typically unimodal with low variance, enabling us to generate a single ground truth aesthetic appeal factor for each image (F_a) by averaging its assigned scores.

To truly understand how the rules of composition affect the ranking system, on a different setting, participants were divided into 3 groups and a subset of 20 randomly-selected images were assigned to each group. Users of each group were asked to select the foreground and specify a region in background, where they wished the foreground object to be placed while preserving the scene semantics, for instance, the boat stays in water. Perspective correction and images were further touched up to remove segmentation artifacts. The ranking exercise was then interchanged between the groups, and a corresponding F_a is obtained per modified image. We observed the following interesting trends in the rank assignment among the images: (1) images with $1 < F_a \leq 2$ received 91% of the votes marked as 1 and 2, and (2) images with $4 < F_a \leq 5$ received 88% of the votes marked as 4 and 5. This indicates

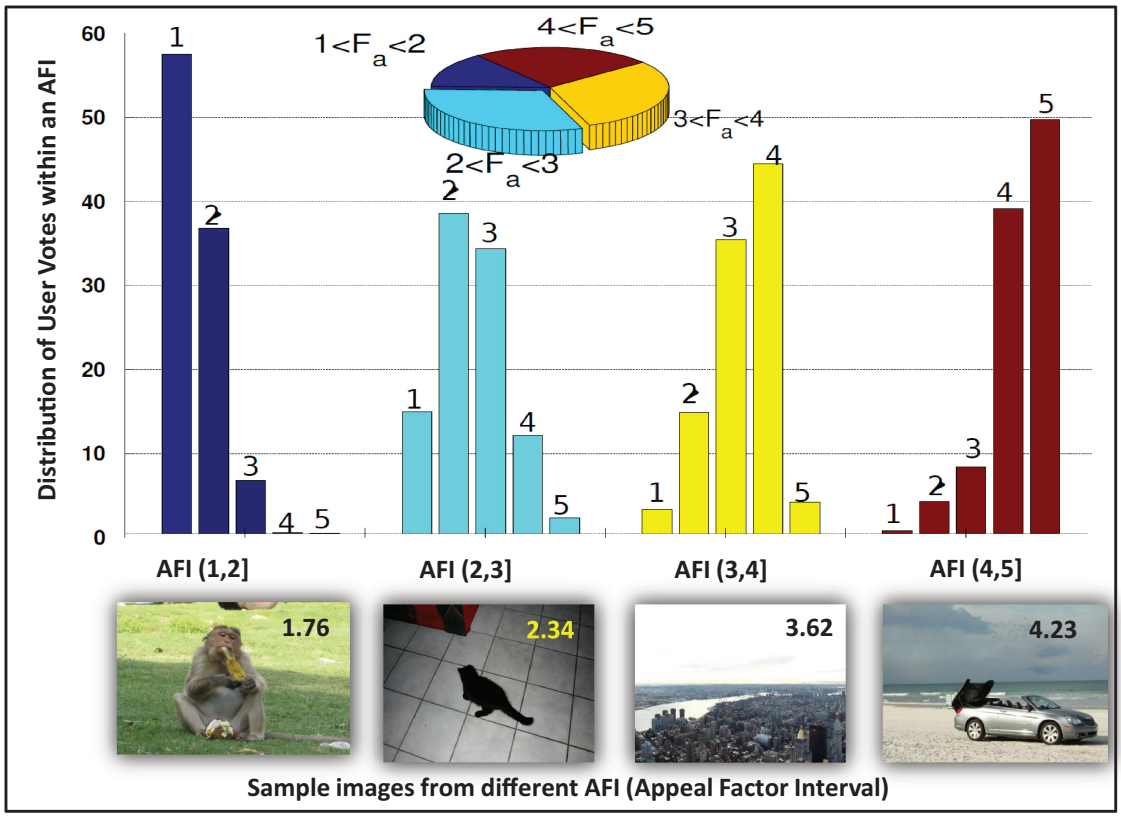


Fig. 5. Summarized information from our user survey. The pie chart shows the distribution of the ground truth aesthetic appeal across various images in our dataset. Each sector in the pie chart corresponds to an appeal factor interval ($(1, 2], \ldots, (4, 5]$). The bar graph shows the distribution of the assigned ranks within each interval: for instance, in the interval $1 < F_a \leq 2$, we observe a large number of images that are ranked 1 by most users. The bottom row of images contain a sample image from each of the four appeal factor intervals with their corresponding appeal factors obtained after combining the user votes.

that the participants are clearly able to distinguish between a well-composed and a poorly-composed image based on the foreground's spatial location in the image frame; these results are detailed in Figure 5.

4.3 Aesthetic Features

In order to formulate photographic quality assessment in the context of a machine learning problem, we need to associate the users' notions of aesthetics to well defined, composition-specific features from an image. To this end, we extract a *relative foreground position* feature for images with single-foreground compositions, and a *visual weight ratio* feature for photographs of seascapes or landscapes. Both of these features are based on elementary rules of photographic composition and are discussed as follows.

Relative foreground position is defined as the normalized Euclidean distance between the foreground's center of mass, also called the *visual attention center*, to each of four symmetric *stress points* in the image frame. This rule can be traced back to painting depicted by renowned artists from the

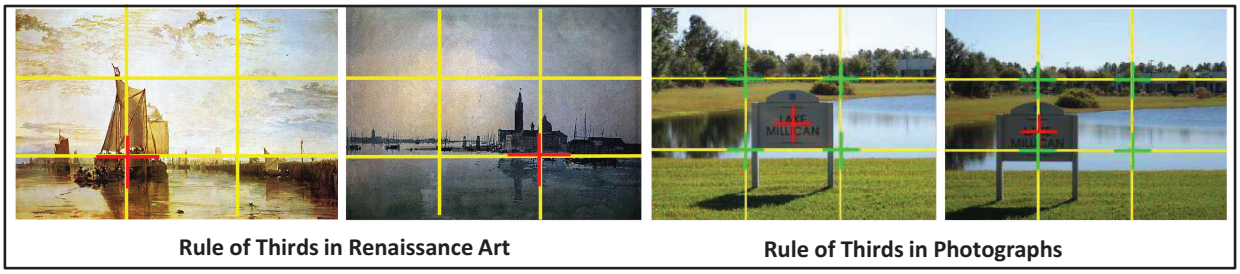


Fig. 6. Rule of thirds in Renaissance painting and modern day photographs: All four images describe the relationships between visual attention center (which is the center of mass of the dominant foreground object) and four stress points (adapted from the rule of thirds). Yellow lines divide the rectangular frame into nine identical rectangles. Each intersection of the yellow lines generates a stress point. This is indicated by green crosshairs in the photographic frames. The red crosshairs in each image mark the foreground object's visual attention center. The third image from the right, shows a photograph taken with the object placed in the middle of the frame. The same scene photographed after aligning the visual attention center close to the stress-point on the bottom left. (Best viewed in color.) [Image of the paintings courtesy of Bruce M. Johnson, http://hoocher.com/Joseph_William_Turner/Joseph_William_Turner.htm])

Renaissance period. An analogy is shown in Figure 6. In photographic literature [Jonas 1976], the stress points are the strongest focal points in a photographic frame (indicated by green cross-hairs in the two rightmost images in Figure 6). In order to attract the viewer's attention to a foreground, the photographer is often advised to adjust the frame in a way so that the foreground's center of mass (red cross-hair) coincides with one of these stress points. The clause, "one of these stress points," is of particular interest in this context, since if the visual attention center is positioned equidistant from all the stress points during the capture, the viewers' attention gets equally divided across these four points. This causes the viewer to lose interest in the photograph, thereby reducing its aesthetic appeal (see third image from right in Figure 6). This observation is also confirmed by our user study where participants tend to rank images with foreground aligned near a stress point higher than those with foreground centered in the frame.

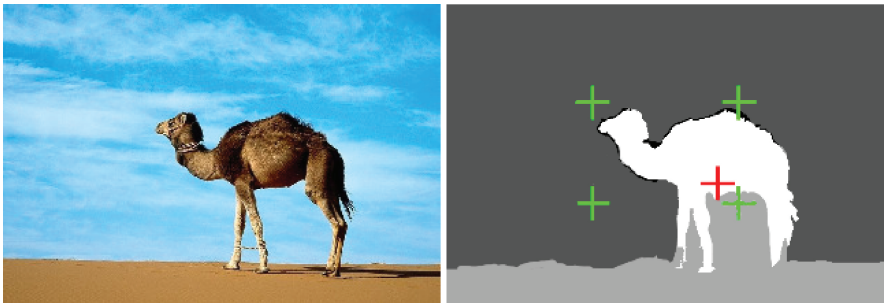
Thus, every photograph containing a single subject composition can be uniquely characterized by a four dimensional feature vector ($\mathbf{F}$):

$$\mathbf{F} = \frac{1}{h \times w} \left[||\mathbf{x}_0 - \mathbf{s}_1||_2, \ldots, ||\mathbf{x}_0 - \mathbf{s}_4||_2 \right], \tag{1}$$

where h, w are the height and width of the image, $\mathbf{x}_0$ is the visual attention center and $\mathbf{s}_i$ are the stress points starting from top left, in clockwise direction. Figure 7 shows two single subject compositions from our dataset, with their respective visual attention center and stress point locations and the corresponding appeal factor of these images, obtained from the user study with the computed $\mathbf{F}$ values. Figures 7 and 8 demonstrate two automatic techniques that we have used throughout this paper for segmenting the foreground from the background and extracting vital semantic information about the scene.

Although the relative foreground location is effective for compositions with one or more dominant foreground objects, it is inapplicable for the class of images in our dataset that consist of landscape or seascape scenes, that do not contain a compact foreground object. For such images, we formulate a second set of features.

Visual weight ratio can be described as the ratio of approximate number of pixels in the sky region, to that in the support region (ground or sea). Compositions conforming to this, like the rule of thirds, are widely seen in Renaissance paintings. For a given landscape or seascape, we estimate the visual



F_a	Relative Foreground Location (F)			
	Top-left	Top-Right	Bottom-Right	Bottom-Left
4.17	0.4381	0.4477	0.0233	0.3935

Fig. 7. Determining visual attention center using segmentation technique exploiting geometric contexts. Here dark-gray pixels denote sky, light-gray denote support, and white pixels belong to the dominant foreground object. Red and green crosshairs indicate the locations of the visual attention center ($\mathbf{x_0}$) and the four stress points ($\mathbf{s_1} \ldots \mathbf{s_4}$) in the frame. Note that the foreground object's outline is more detailed in this case, compared to the saliency based technique illustrated in Figure 8, which makes the former a better fit for the recomposition technique, discussed later. The adjacent table shows a mapping between the aesthetic appeal factor (F_a) and the relative foreground location feature ($\mathbf{F}$), extracted from these two images. The values in the second to fifth columns can be interpreted as the relative Euclidean distances between the visual attention center ($\mathbf{x}_0$) and the four stress-points ($\mathbf{s_1} \ldots \mathbf{s_4}$), normalized against the width and height of the image frame.[Image courtesy of Brooks Walker, http://www.spoki.lv/foto-izlases/Kaut-kas-no-national-geographic/12682/1/2]

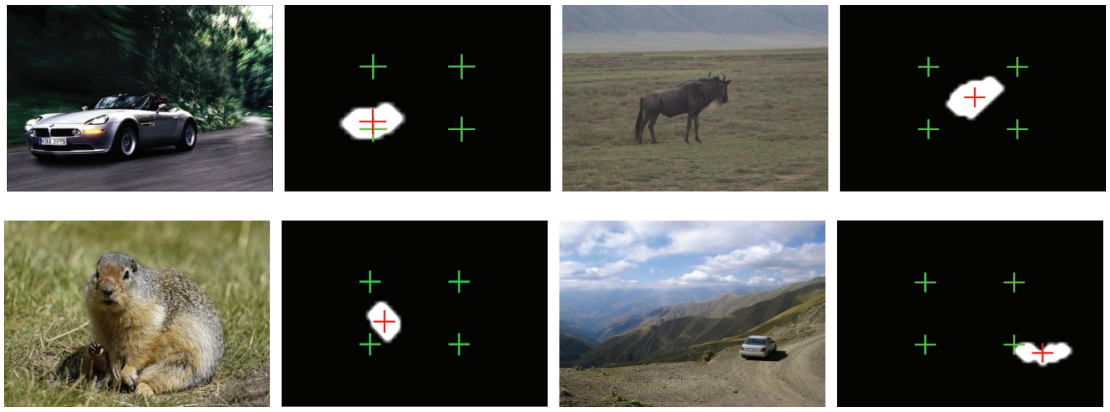


Fig. 8. Saliency based detection of visual attention center. Each row shows two pairs of input and output images, the images in black background rows show the output of a saliency algorithm. Dominant foreground region is shown as a white blob in a black background. Similar to Figure 7, the visual attention center and the four stress points are shown in red and green crosshairs respectively. Note this technique by itself does not provide us with any scene information.

weights in the sky region by the automatic semantic segmentation technique discussed in the beginning of this section. Our tool allows the user to interactively adjust the detected horizon line. The user also has his liberty to etch out the semantic regions in a scene, in case the automatic semantic segmentation results are not satisfactory. Although in this context we assume that the horizon is always a line parallel to the x-axis of the image, we later show in Section 5.1 that we facilitate correction of tilts to enforce that the horizon is always parallel to the x-axis.

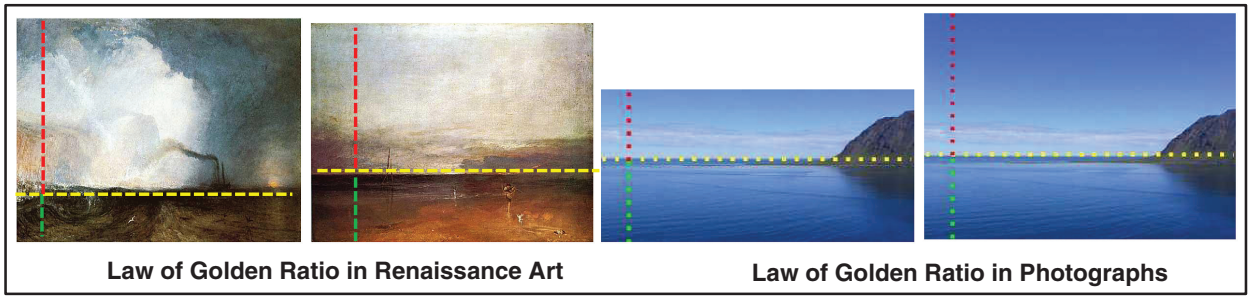
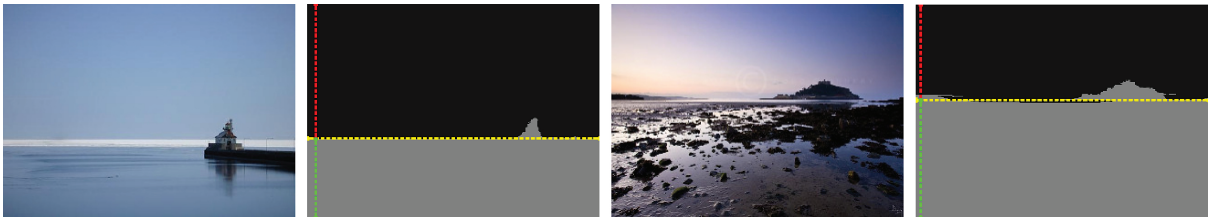


Fig. 9. Quantifying visual weight balance: the yellow dotted line marks horizon, the red dotted line marks the vertical extent of sky (Y_k), and the green dotted line marks the vertical extent of the sea (Y_g). The rightmost image shows a composition with ideal combination of visual weights. The image to the left, shows a cropped version of the same composition with altered visual weights. The latter is rated as more visually appealing. [Image of the paintings courtesy of Bruce M. Johnson, http://hoocher.com/Joseph_William_Turner/Joseph_William_Turner.htm]



F_a	Visual Weight Deviations ($\mathbf{W}$)	
	$\lvert\phi - \frac{Y_g}{Y_k}\rvert$	$\lvert-\phi \frac{Y_k}{Y_k+Y_g}\rvert$
4.58	0.099	0.012
3.23	1.112	0.976

Fig. 10. Visual weight measures for sky and support regions: horizon (yellow dotted line) and the respective vertical extents (red and green dotted lines) are shown for each image. Quantitative interpretation of these extents are provided in the adjacent table. The leftmost column indicates the appeal factor of the two images while the other two columns indicate the visual weight deviations from the Golden Ratio.

The idea behind visual weights can be illustrated with the help of Figure 9. The leftmost images are compositions from Renaissance paintings while the third and forth images from the left are actual photographs of the same scene. In all the images, the horizon separates the frame into two rough rectangles. The ratios between the areas of these rectangles should be close to the golden ratio [Livio 2002] for a better appeal, that is,

$$\frac{Y_g}{Y_k} = \frac{Y_k}{Y_k + Y_g} = \phi, \quad Y_k > Y_g, \tag{2}$$

where Y_k (red dotted line), Y_g (green dotted line) denote the vertical extents of sky and support regions respectively in Figure 9 and ϕ is the golden ratio. In order to maintain the aesthetic balance, these ratios should be equal to the golden ratio (ϕ),which is approximately equal to 1.61803. For the rightmost image in Figure 9, these these ratios are observed to be 1.6011 and 1.5934 ($\approx \phi$), while for the the third image in Figure 9, the same numbers are 0.4533 and 0.6743, which makes the former more appealing of the two. A mapping of the visual weights to the user assigned aesthetic appeal of two sample images from our training dataset is shown in Figure 10.

We make a reasonable assumption that the photographic frame is approximately aligned with the horizon so that Y_k and Y_g could be estimated by averaging the vertical extents of the pixels belonging to sky and support regions, respectively. For images with substantial tilt, tilt correction is performed in the preprocessing stage. The deviations of the individual ratios $\frac{Y_g}{Y_k}$ and $\frac{Y_k}{(Y_k+Y_g)}$ from the Golden Ratio (ϕ) form the aesthetic feature ($\mathbf{W}$) for photographs of seascapes or landscapes. Formally,

$$\mathbf{W} = \left[\left| \phi - \frac{Y_g}{Y_k} \right|, \left| \phi - \frac{Y_k}{Y_k + Y_g} \right| \right]. \tag{3}$$

The two high-level features discussed previously are clearly not the only ones that can capture an image's aesthetic appeal. They were chosen because they can be reliably quantified using existing techniques and address typical photographs found in Internet photo collections. Other metrics from the photography literature are either too abstract, demanding a sophisticated understanding of the image scene that is beyond current computer vision algorithms, or would apply to only a relatively small subset of photographs.

4.4 Learning and Prediction

The aesthetic appeal for the two different types of photographic composition that we have addressed here can be associated with the features extracted using two smooth functions defined as:

$$f_{rf}(F_a) : R^4 \rightarrow R, R \in \mathbf{F}, \tag{4}$$

$$f_{vw}(F_a) : R^2 \rightarrow R, R \in \mathbf{W}, \tag{5}$$

based on Equations (1) and (3). We use two independent, soft-margin support vector regressors implemented using Joachims [1999] to learn these nonlinear mappings. We employ a coarse grid search with the SVR's error parameter values (C) from 0.1, 1, 10, and tube-width values (ϵ) from 0.01, 0.1, 1, 10 on an RBF kernel with σ values from 0.5, 1, 2. We select 150 random images from either composition class for training and use the rest for testing. The best prediction accuracy of $87.3 \pm 3\%$ for photographs with single foregrounds is reported for $\sigma = 2$, $C = 0.1$, $\epsilon = 1$. The same number for the latter composition category is reported to be $96.1 \pm 2\%$. A detailed quantitative analysis is provided in the results section.

5. ENHANCEMENT ENGINE

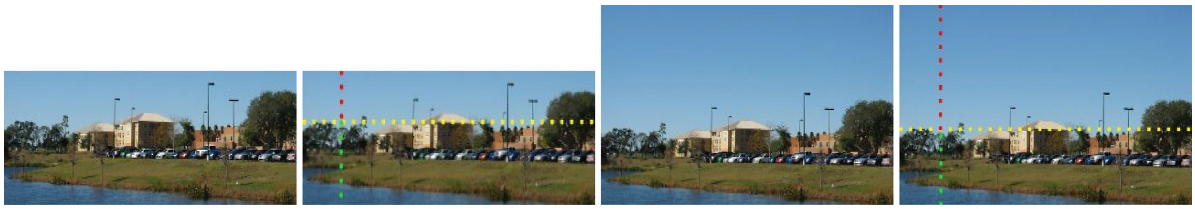
Our recomposition technique is built upon inputs from the same aesthetic features that used to evaluate a given composition. We introduce a pipeline of algorithms as discussed towards the beginning of the article in Figure 3. As a preprocessing stage, we eliminate any tilt observed in a given image in order to align the photographic frame with the horizon. This is discussed in Section 5.1 and is required by the following stages in the enhancement engine. The first stage focuses on increasing the appeal factor of landscape and seascape images by better balancing the visual weights of the sky and support regions. This is discussed in Section 5.2. This technique can be directly applied to any image that have a clearly demarcated horizon in conjunction with the next stage which involves optimally placing a foreground object in the image so that the composition conforms to aesthetic rule of thirds, while maintaining the integrity of the scene.

5.1 Tilt Correction

In order to rectify image tilt, the user interactively selects two points on the horizon that are joined to form a straight line. This line is our tilted horizon. Next we draw a straight line passing through one of

(a) (b) (c) (d)

Fig. 11. Rectifying image tilt before balancing visual weights: (a), (c) Original images with tilted horizon, (b), (d) Respective images after tilt correction.



(a) (b) (c) (d)

Fig. 12. Altering a composition to balance visual weights: (a) Original image. (b) Corresponding image showing the distribution of visual weights. (c) The vertical extent of sky increased using our method to balance the distribution of visual weights, improving the overall aesthetic appeal of the image; (d) Modified distribution of the visual weights.

the user selected points parallel to the x-axis of the rectangular image frame. We refer to this line as the image horizon. Using the angle between these two lines, we can easily compute a rotation transform that can be applied to rectify the tilt in the image. Two typical compositions with substantial tilt are shown in Figure 11(a), and Figure 11(c). The respective tilt corrected results are shown in Figure 11(b) and Figure 11(d). We automatically crop the transformed image to eliminate the artifacts generated near its image borders due to the rotation.

5.2 Visual Weight Equalization

Let us assume that a horizontal line divides our image in the ratio $\frac{Y_k}{Y_g}$. This can be visualized in Figure 12, where the horizon is represented by the yellow dotted lines. A fixed-step Y_k expansion or contraction strategy can be applied here to solve Equation (5), which leads to the optimal combination of visual weights that maximizes the appeal factor.

Since this is relatively less complex than solving for the optimal foreground placement location, we resort to a simpler technique by assuming the following holds good at the optimal solution:

$$\frac{Y_k}{Y_g} = k\frac{Y_g}{Y_k + Y_g}, \quad k > 0. \tag{6}$$

Let h be the vertical extent that Y_k must be increased so that:

$$\frac{Y_k + h}{Y_g} = \frac{Y_g}{(Y_k + h) + Y_g}. \tag{7}$$

(a) (b) (c) (d)

Fig. 13. Combining recomposition algorithms: (a) Original image, (b) Tilt rectified background image, (c) Visual weight balancing applied on tilt rectified image, and (d) Optimal object placement algorithm applied on the visually balanced image followed by adequate scaling and inpainting.

With a couple of algebraic substitutions in Equation (7) from Equation (6), we obtain a quadratic equation in h, which can be easily solved for two values of h. A positive value of h indicates an increase of Y_k by h, while a negative value of h means decrease of Y_g by h, leading to an increase or decrease in the overall image height. In order to increase the height of the image, we are required to in-paint the newly-added region with information available from neighboring pixels. Decreasing the height is simply performed by cropping the image appropriately. For inpainting, we employ the straightforward patch-based region filling algorithm proposed by Zhang et al. [2004]. We limit the search for target patches in 20×20 neighborhood of the source patch. For most of the images in our dataset, we achieve aesthetically pleasing results with fewer than 60 iterations of a graphcut-based patch updating mechanism discussed in Zhang et al. [2004]. However, the algorithm frequently introduces minor artifacts into the background of our recomposed images, that require interactive retouching. Additional results are shown in Figure 22.

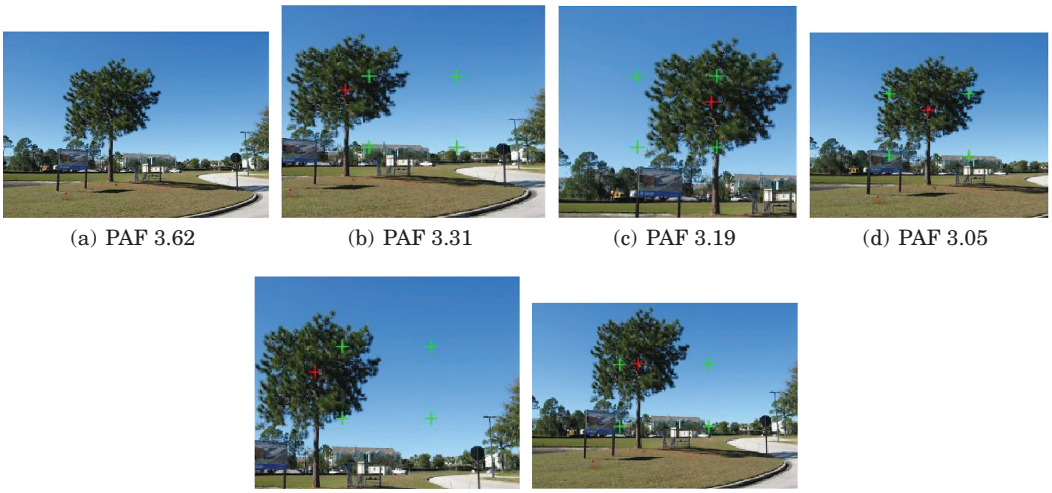
The algorithms discussed in the previous sections can be applied in a sequential order on certain images to increase their aesthetic appeal. An example is shown in Figure 13.

5.3 Optimal Object Placement

The problem of spatial recompositing is closely related to the simpler task of optimally cropping a given photograph in order to enhance its visual appeal as studied by [Nishiyama et al. 2009]. Since the locations of the stress points are determined entirely by the frame dimensions, one can crop a photograph to better align the dominant object with a given stress point, as shown in Figure 14.

Unfortunately, while this analogy prescribes a straightforward solution to the problem of optimal foreground alignment, it is unsatisfactory in two key respects. First, cropping reduces the size of the image frame and can alter its aspect ratio. Second, and more importantly, cropping can lead to the loss of valuable image information, such as key aesthetic features in the background. This motivates us to attempt a more ambitious goal: moving the foreground object in the image frame to a better location without compromising the semantics of the scene. In the context of Figure 15, we seek to move the foreground object (tree) in such a manner that the predicted appeal factor after the relocation increases while keeping the tree in contact with its support in the background.

Recall $\mathbf{x_0}$ as the location of the current *visual attention center* (the foreground object's centroid in image coordinates), we define the support neighborhood for the foreground as ψ_w. In other words, these are the set of pixels that lie within $w \times w$ neighborhood of the boundary of the foreground. With a slight abuse of notation, let $\psi_w(\mathbf{x_0})$ denote the set of pixels forming the *support neighborhood* at



(a) PAF 3.62 (b) PAF 3.31 (c) PAF 3.19 (d) PAF 3.05

(e) PAF 3.48 (f) PAF 3.69

Fig. 14. Illustrating an analogy between ideal positioning of the subject and optimally cropping the photograph: (a) original image; (c)–(e) cropped samples of the original image that move the visual attention center (centroid of the tree, denoted by a red crosshair) towards/away from the stress points (green crosshairs); (f) a near-optimal crop that aligns the visual attention center near the top-left stress point. For every crop, the respective appeal factor is determined using the relative foreground location feature based regressor.

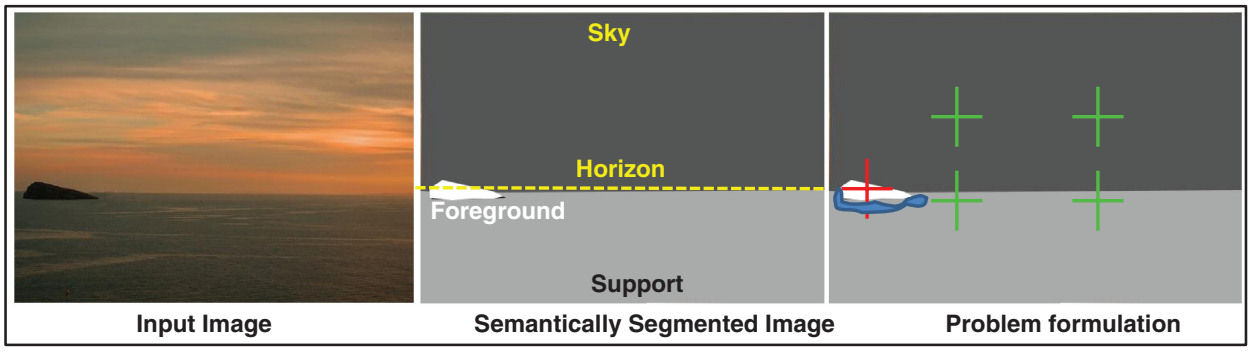


Fig. 15. Formulating the optimal object placement problem: Original image; dominant foreground object, sky and support regions are represented using white, dark gray, and light gray pixels respectively; Blue pixels are special cases of support pixels in the foreground object's neighborhood (ψ_w); four green crosshairs mark the stress points; red crosshair marks the visual attention center ($\mathbf{x_0}$).

the object's original location and $\psi_w(\mathbf{x})$ to be those pixels that would form the support neighborhood were the object mask to be centered at $\mathbf{x}$ rather than $\mathbf{x_0}$ at a single iteration. Clearly, the shape of the support neighborhood is constant for any $\mathbf{x}$, but the intensity values of the underlying pixels (in each of the three channels, assuming an RGB colorspace) from the background would change. Now, we express the problem of relocating the object to an aesthetically favorable location $\hat{\mathbf{x}}$ as the following optimization problem:

$$\arg\max_{\mathbf{x}} f_{rf}(F_a) \quad \text{s.t. } \lambda(\mathbf{x}, \mathbf{x_0}) < \delta, \tag{8}$$

(a) PAF 2.32 (b) PAF 3.81 (c) PAF 3.17 (d) PAF 3.78 (e) PAF 4.39

Fig. 16. Intermediate results from spatial recomposition with corresponding predicted appeal factor computed by the rule of thirds based regressor: (a) original image; (b)–(d) potential solutions for relocating foreground object; (e) optimal location. Note that these results do not include rescaling the object in accordance with perspective as described in Section 5.3.1.

where δ is a human-specified real-valued number which enforces how closely the support regions must match, and $\lambda(\mathbf{x}, \mathbf{x_0})$ is a smoothness term computed over the pixel intensities and gradients in the spatial neighborhoods of $\mathbf{x}$, $\mathbf{x_0}$ as:

$$\lambda(\mathbf{x}, \mathbf{x_0}) = S_I + \beta S_\nabla. \tag{9}$$

Here β is a regularization parameter, usually set to a high value (∞) for regions with large texture variations, S_I and S_∇ are the intensity and gradient components of the smoothness term respectively, calculated as:

$$S_I = \sum_{\psi_w \forall \{R,G,B\}} ||I(\psi_w(\mathbf{x})) - I(\psi_w(\mathbf{x_0}))||_1, \tag{10}$$

$$S_\nabla = \sum_{\psi_w \forall \{R,G,B\}} ||\nabla(\psi_w(\mathbf{x})) - \nabla(\psi_w(\mathbf{x_0}))||_1. \tag{11}$$

The solution to Equation (8) gives us the new location for the visual attention center of the foreground object ($\hat{\mathbf{x}}$). We obtain $\hat{\mathbf{x}}$ by optimizing using standard techniques. Figure 16 shows some intermediate outputs from our algorithm during the optimization process. We observe that the location of the horse shifts from frame to frame. In the best result, the location of the horse is well aligned with a stress point and the support neighborhood is highly consistent with that of the original image. We explicitly set the search window to a homogeneous grass-covered region, for a faster convergence.

We discuss whether (and how) to scale the horse to correct for perspective, and how to inpaint the hole left at its original location later in the paper. Given the small size of this optimization problem, we use an exhaustive search with a user-specified quantization size to optimize Equation (8) as this guarantees a globally optimal solution. Furthermore, we reduce the complexity of the search from $O(h \times w)$ to $O((h-l) \times (w-m))$ where l, m are dimensions of the region that is semantically least likely to contain the foreground after recomposition. A detailed qualitative and quantitative analysis of the recomposition technique is provided in Section 5.3.1.

5.3.1 *Perspective Correction.* Simply translating the foreground object in the scene is insufficient for photo-realistic recomposition. This is because moving an object vertically in the image changes the depth at which it is perceived in the scene. For instance, an object on the ground should shrink as it translates up in the image by a factor that depends on imaging characteristics such as the focal length and tilt of the camera. Thus, spatial recomposition must correctly rescale the object to maintain photorealism.

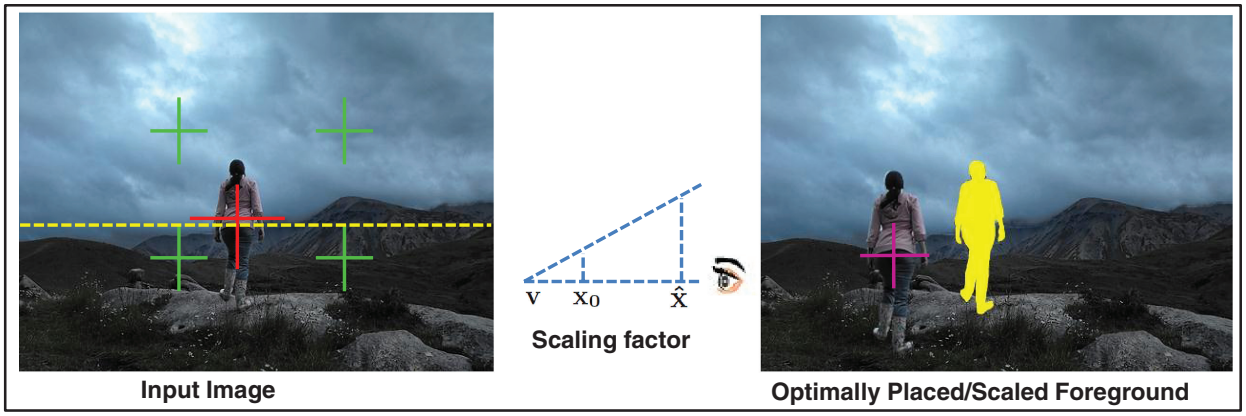


Fig. 17. Illustration of perspective correction: The image on the left shows the original image overlaid with the location of the horizon, the four stress points shown in green crosshairs, and the visual attention center ($\mathbf{x_0}$) in red cross-hair. The schematic diagram in the middle provides an intuition to compute the scaling factor based on the vanishing point. Finally the image with the scaled foreground overlaid with a purple cross hair depicting the optimal location of the foreground ($\hat{\mathbf{x}}$). The cutout of the foreground object is marked in yellow.

Fortunately, we can employ methods that automatically estimate the location of the horizon in the image [Hoiem et al. 2007] to determine the correct size of the foreground object at its new location using the following straightforward equation:

$$v_x = \frac{D_x}{D_y}(v_y - y_2) + x_2, \tag{12}$$

where $\mathbf{v} = (v_x, v_y)$ is the vanishing point [Hartley and Zisserman 2004] that is, the point of intersection between the horizontal line $y = v_y$ and the line through the original object location $\mathbf{x_0}$ and its modified location $\hat{\mathbf{x}}$. D_x/D_y is the slope of this line and x_2, y_2 are the components of $\hat{\mathbf{x}}$. The scaling factor is computed as:

$$f_s = \frac{||\mathbf{v}, \mathbf{x_0}||_2}{||\mathbf{v}, \hat{\mathbf{x}}||_2}. \tag{13}$$

For images where the vanishing line information cannot be reliably determined, we simply keep the size of the object constant (equivalent to orthographic projection). We show our results in Figure 17, where the image on the right shows a slight increase in size as the foreground object moves towards the viewer in the image frame. A fast bicubic interpolation algorithm is applied to perform the scaling operation. More results are shown in Figure 20.

5.3.2 *Multiple Subjects.* The proposed idea of spatial recomposition using optimal object placement can be extended to enhance compositions featuring multiple subjects. This is accomplished in an n-way incremental fashion, where n is the number of subjects in foreground that are required to be relocated. It is very difficult to model the aesthetics of compositions that span more than three foreground ($n \leq 3$) subjects using simple guidelines such as the rule of thirds [Jonas 1976]. Besides as the number of subjects increase in a photographic frame, the complexity of their spatial relationships also increases. In this section, we first demonstrate recomposition with two subjects in foreground and then generalize it for more.

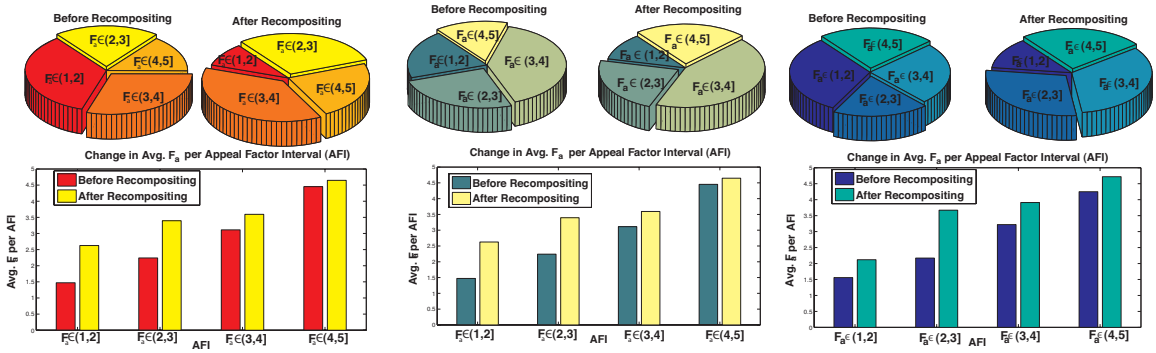
Fig. 18. Recomposition with multiple subjects: (a) Original image containing two subjects (ball and bird), (b) Image with respective visual attention centers (red crosshairs) shown for reference, (c) Result of recomposition with selecting the ball as primary subject and removing the bird from the image. (d) Recompositing with the second subject(bird), the first subject (ball) becomes part of the background. This avoids situations where subjects can overlap. (e)–(f) Additional recomposition results with two subjects (person with kayak and chair), and (g)–(h) with three subjects (two volleyball players and a volleyball).

Suppose that we have an image with two foreground subjects S_a and S_b. We begin by removing S_b from the composition and filling the hole created in its location using an inpainting algorithm [Zhang et al. 2004] (detailed later). We optimize Equation (8) for S_a treating the image as a single subject composition. Next, S_a is treated as part of the background and we optimize Equation (8) for S_b. The maximum appeal factor is recorded for this configuration. We then repeat the same process with the order of foreground object selection reversed. Hence, for a two subject composition we obtain two independent optimal appeal factors. The recomposition that results in the maximum appeal factor is retained as a globally optimal solution. Figure 18 shows some interesting results of recomposition with multiple subjects. Additional results are shown in the experimental results section (Figure 21).

6. EXPERIMENTAL RESULTS

We performed an extensive qualitative and quantitative evaluation of the proposed methods, summarized as follows. We apply the proposed recomposition techniques separately to 200 images taken from both categories(single object compositions and sea/landscapes) of the dataset. This is facilitated by a graphical tool where a user is interactively asked to label regions sky, support or the foreground object using closed polygons. An automatic segmentation option is also provided which can be used for relatively less complex scenes, for example scenes without shadows, reflection etc. Once the user is satisfied with the segmentation process, he/she chooses which algorithm to apply. Depending on the algorithm selected, the tool employs either of the two techniques discussed in Section 5.3 or Section 5.2 or a combination of both for single as well multiple subjects based composition in Section 5.3.2.

Of the 200 images in the single subject composition category, 38 have appeal factors in the interval $(1, 2]$, 49 in $(2, 3]$, 75 in $(3, 4]$, and the rest are in the last interval $(4, 5]$. The recomposited images are



(a) (b) (c)

Fig. 19. Quantitative results on images recomposited using (a) Visual weight equalization (refer to Section 5.2) technique. In both pie-charts, each sector represents the fraction of images whose respective appeal factor lie in one of the four discrete intervals $((1, 2], (2, 3], (3, 4], (4, 5])$. Recompositing shows definite improvement in the lower two intervals as their respective sectors shrink in the right pie-chart. The bar-chart in the bottom shows the net improvement of appeal factors pertaining to each intervals after recompositing. (b) optimal object placement (refer to Section 5.3) technique, and (c) combination of both algorithms, on images with multiple foreground (refer to Section 5.3.2). We observe a similar trend as seen in Figure 19(a) and Figure 19(b).

then evaluated by users in the same way as discussed in Section 4.2. We observe a clear increase in aesthetic appeal of images whose F_a values were in the $(1, 2]$ and $(2, 3]$ intervals as sectors corresponding to these intervals shrink in the rightward pie chart in Figure 19(b). The increased area of the sector corresponding to the interval $(3, 4]$ in the same pie chart show in favor of the argument that some images from the lower intervals have moved up, after recompositing. Since the aggregated statistics shown in the pie-chart do not provide insight on how individual images could have been affected as a result of the process, we also plot the average appeal factors of images in each interval, before and after recompositing.

A similar experiment is performed for the land/sea scape images. In this case, we begin with 82 images whose appeal factors are in the interval $(1, 2]$, 86 in $(2, 3]$, 21 in $(3, 4]$, and the rest in $(4, 5]$. We see a similar trend as observed in Figure 19(b) in this setting (Figure 19(a)) as well. The bars corresponding to the interval $(4, 5]$ indicate that there is little scope for improvement for images that are already aesthetically appealing.

Some qualitative results obtained after recomposition are given in Figures 20, 21, and 22. Note how the scales are adjusted for foregrounds in some of the images (person, cow, building, boat) with inputs from user about the respective scenes. The bottom two rows show some results after applying the visual weights based recomposition. Here nonsky region is cropped optimally to increase the visual appeal, these images show results of sky-region augmentation to increase the appeal.

7. CONCLUSION

We have introduced a new multimedia application that enables users to automatically assess the aesthetic quality of a photograph using geometric rules of composition, and then to make an informed decision on how to improve the photograph using spatial recomposition. Rather than prescribing a fully-automated solution, we allow user-guided object segmentation and inpainting to ensure that the final photograph matches the user's criteria. Our approach achieves 86% accuracy in predicting the attractiveness of unrated images, when compared to their respective human rankings. Additionally, 73%

Fig. 20. Results of spatial recomposition using optimal object placement on a subset of images from our dataset that depict single subject compositions (Success): Each pair of images has the original image on the left and its recomposited counterpart on the right.

of the images recomposited using our tool are ranked more attractive than their original counterparts by human raters.

In future work, we plan to replace the resizing operations currently used in recomposition by a more context-aware resizing algorithm [Avidan and Shamir 2007]. We also plan to fuse information available from multiple compositions of the same scene to create a more aesthetically pleasing photograph. In our tool, we intend to use a more sophisticated segmentation algorithm with minimal intervention that

Fig. 21. Additional results of spatial recomposition on multiple foreground objects.

Fig. 22. Results of spatial recomposition using visual weight balancing on landscape or seascape images. The visual weights are optimally altered to make the images more visually appealing.

can generate additional training data for a robust aesthetic model that could be applied to improving Internet image search. In addition, we would like to explore how our enhancement technique could be applied synergistically with low-level image editing techniques [Liu et al. 2008], while preserving the semantic essence of the scene.

REFERENCES

AVIDAN, S. AND SHAMIR, A. 2007. Seam carving for content-aware image resizing. In *Proceedings of the ACM SIGGRAPH International Conference on Computer Graphics and Interactive Techniques*.

BOUTELL, M. AND LUO, J. 2004. Bayesian fusion of camera metadata cues in semantic scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

CHO, T. S., BUTMAN, M., AVIDAN, S., AND FREEMAN, W. T. 2008. The patch transform and its applications to image editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision*.

DATTA, R., LI, J., AND WANG, J. Z. 2007. Learning the consensus on visual quality for next-generation image management. In *Proceedings of the ACM Multimedia Conference*.

HARTLEY, R. AND ZISSERMAN, A. 2004. *Multiple View Geometry in Computer Vision*, 2nd Ed. Cambridge University Press.

HOIEM, D., EFROS, A., AND HEBERT, M. 2007. Recovering surface layout from an image. *Int. J.Comput Vis. 75*, 1.

JOACHIMS, T. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods: Support Vector Learning*. MIT Press.

JONAS, P. 1976. *Photographic Composition Simplified*. Amphoto Publishers. 2, 6, 15.

KE, Y., TANG, X., AND JING, F. 2006. The design of high-level features for photo quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

LEYVAND, T., COHEN-OR, D., DROR, G., AND LISCHINSKI, D. 2008. Data-driven enhancement of facial attractiveness. In *Proceedings of the ACM SIGGRAPH International Conference on Computer Graphics and Interactive Techniques*.

LIU, C., SZELISKI, R., KANG, S. B., ZITNICK, C. L., AND FREEMAN, W. T. 2008. Automatic estimation and removal of noise from a single image. *IEEE Trans. Patt. Anal. Mach. Intel. 30*, 299–314.

LIVIO, M. 2002. The golden ratio and aesthetics. *Plus Mag. Living Math*.

LUO, Y. AND TANG, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the European Conference on Computer Vision*.

MANSOOR, A., HAIDER, M., MIAN, A., AND KHAN, S. 2009. A hybrid image quality measure for automatic image quality assessment. In *Proceedings of the Scandinavian Conference on Image Analysis*.

NISHIYAMA, M., OKABE, T., SATO, Y., AND SATO, I. 2009. Sensation-based photo cropping. In *Proceedings of the ACM Multimedia Conference*. 669–672.

SUN, X., YAO, H., JI, R., AND LIU, S. 2009. Photo assessment based on computational visual attention model. In *Proceedings of the ACM Multimedia Conference*. 541–544.

VENKATA, N. D., KITE, T. D., GEISLER, W. S., EVANS, B. L., AND BOVIK, A. C. 2000. Image quality assessment based on a degradation model. *IEEE Trans. Image Process. 9*, 2.

WALTHER, D. AND KOCH, C. 2006. Modeling attention to salient proto-objects. *Neural Networks* 19, 4.

YOU, J., PERKIS, A., HANNUKSELA, M., AND GABBOUJ, M. 2009. Perceptual quality assessment based on visual attention analysis. In *Proceedings of the ACM Multimedia Conference*.

ZHANG, Y., XIAO, J., AND SHAH, M. 2004. Region completion in single image. In *Proceedings of EUROGRAPHICS*.

Received March 2011; revised August 2011; accepted September 2011