SALIENCY-BASED IMAGE ENHANCEMENT

LAI-KUAN, WONG

NATIONAL UNIVERSITY OF SINGAPORE

2013

SALIENCY-BASED IMAGE ENHANCEMENT

LAI-KUAN, WONG
M.Sc., National University of Singapore

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE

2013

To my husband, Tom,
who gives me the wings to fly.

Acknowledgement

First and foremost, I would like to express my heartfelt gratitude and appreciation to my supervisor, Low Kok Lim. He has offered me invaluable guidance, and constructive ideas throughout my graduate studies. He also contributed his invaluable time and effort to carefully review all research papers, which indirectly, taught me the art of writing a good and precise research paper. It was indeed great working with him and I will always be thankful.

I am very grateful to Terrence Sim and Michael S. Brown who have taught me the fundamentals of Computer Vision and Computational Photography respectively. Knowledge obtained from these two important fields of study helped me to build a strong foundation for my research work. In addition, I also thank them for their valuable comments and suggestions on my GRP and thesis proposal.

My sincere gratitude and respect to Leow Wee Kheng, who has been a great inspiration to me, both as a dedicated lecturer and a researcher, since the beginning of my graduate studies. From his course Multimedia Analysis, I have learnt the invaluable lessons on defining research problems mathematically and solving problems systematically, skills and knowledge that were undoubtedly proven useful throughout the course of my graduate studies.

I would like to thank Tan Tiow Seng and Huang Zhiyong for their precious comments and suggestions on my research during the weekly meeting of G3 Lab. Special thanks to Vlad Hosu, my ex-project partner who then became a good friend, for showing me new and creative ways of tackling research problems. Not forgetting to thank all my fellow lab-mates who offered me great company and assistance in many ways. They have enriched my life in NUS, making it more enjoyable and fun.

I would like to express my heartfelt appreciation to my ever-supportive family and friends. My deepest gratitude goes to parents, especially my mom for her unconditional love, care and support. I thank all my sisters and cousin Yoke Mun who are always here for me, offering support and encouragement. I am also truly blessed to have some friends who never fail to offer spiritual support and always ready to lend a helping hand. Special thanks to Ming Kee, Soh Hong, Thiam Chiew and Hooi Mien for making my life more meaningful, interesting and enjoyable.

Last but not least, I thank my husband, Tom for letting me fly and never stop me from pursuing my dream. Without his love, understanding, continuous encouragement and unwavering support, I would not have reach this far.

Abstract

A photograph that has visually dominant photo subjects in general induces stronger aesthetic interest. Prolonged searching for the subjects can reduce the satisfaction of viewing the photograph leading to decrease of aesthetics experience. It is essential to make subjects of interest dominant so that viewers' attention is directed to what a photographer wants them to see. Motivated by the importance of visual dominance in influencing aesthetics, and the lack of research in enhancing visual dominance as a means to improve image aesthetics, in this thesis, we adopt a saliency-based approach for image aesthetics evaluation and enhancement.

The contributions of this thesis are threefold. First, we present the **saliency-enhanced approach for aesthetics class and score prediction**. Our aesthetics class prediction model produces higher classification accuracy compared to state of art approaches. Our score prediction model is proven to be effective in inferring relative aesthetics score of similar images to guide image enhancement. Next, we introduce **saliency retargeting**, a novel low-level image enhancement approach aimed to enhance image aesthetics by redirecting viewers' attention to the important subjects of the scene. This approach applied non-uniform modification to three low-level image features; intensity, color and sharpness that directly correspond to features used in Itti-Koch visual saliency model. Our score prediction model is used to drive the saliency retargeting algorithm to return the maximally-aesthetics version as the result. Finally, another significant contribution of this thesis is **tearable image warping**, a variant of image warping, that can support scene-consistent image recomposition and image retargeting. Capitalizing on the idea that only part of an object is connected to its physical environment, the tearable image warping algorithm preserves semantic connectedness when necessary and allows an object in an image to be partially detached from its background. For *image retargeting*, this approach significantly reduced distortion compared to pure image warping and is able to preserve semantic connectedness such as shadow, which oftentimes can be violated in results of scene carving. For *image recomposition*, our approach can produce an effect analogous to change of viewpoint without semantics violation, making it a powerful recomposition tool. With this capability, we can effectively apply geometric transformation to enhance the visual dominance of the photo subject and other aesthetics elements. Empirical evaluations with human subjects demonstrate the effectiveness of both the saliency retargeting and tearable image warping algorithms in enhancing image aesthetics.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A great photograph demands an object or point of interest as its main image. Everything peripheral must centre around this key focal point.

Paul Summer

The primary interest of an ideal photograph is in telling one what the subjects looked like at a particular time; a photograph is a means to the end of seeing its subjects. Consequently, substantial aesthetic interest in the photograph is derivative from an aesthetic interest in the subjects (Scruton, 1983). It is therefore essential to make subjects of interest dominant so that viewers' attention is directed to what a photographer wants them to see. Prolonged searching for the subjects of a photograph can reduce the satisfaction of viewing the photograph leading to decrease of aesthetics experience. This explains why many of the photographic

composition rules such as leading simplicity, framing, fill the frame, low depth of field (DOF) lines/S-curves, are targeted to increase the dominance of the main subjects. Figure 1.1 shows some examples that follow these rules. Apart from dominance of subjects, there are other aesthetics elements such as balance, depth and perspective, and geometrical elements that can make a photograph more interesting. These aesthetics elements can be enhanced by following a set of photographic composition rules. For example, balance can be achieved by ensuring visual balance and horizon balance as well as adhering to rules of third.

Figure 1.1. Photographs following different rules to enhance the dominance of the photo subject, (a) fill the frame (Photo courtesy of Jim Crotty), (b) simple and plain background, (c) framing, and (d) leading lines.

Many existing automatic image enhancement methods such as contrast, color or edge/texture enhancement mainly focus on altering global features or low-level local features. Rarely subjects of a photograph are considered in the process of enhancement. Only recently, with the emergence of numerous visual attention models that simulate the human visual system to identify regions of interests (ROIs) in an image, researchers start to look into saliency-based image enhancement. Su et al. (2005) and Gasparini et al. (2007) attempted to enhance the saliency of the photo subject by performing selective de-emphasizing of texture variations and selective

edge enhancement respectively. As both of these approaches are not aesthetically-driven, although they managed to make the subject stands out more from the background, resulting images are not necessarily aesthetically more pleasing. Bae et al. (2006) and Barnajee et al. (2007) achieved more success in enhancing image aesthetics by magnifying the blurriness of image content not-in-focus to simulate low depth of field effect, a photographic technique intended to increase the salience of the photo subject. To our best knowledge, approaches to modify intensity or color contrast between subject and background or a unified approach that enhance multiple low-level features to make a subject more dominant are non-existent.

Apart from modifying the low-level image features, the aesthetics of a photograph can also be enhanced by modifying its spatial composition based on photographic rules. Research on automatic image recomposition is still in its infancy stage. Barnajee (2007) and Kao et al. (2008) attempted to enhance image aesthetics by modifying photographs to conform to selected photographic rules such as rule-of-thirds to bring out the photo subject. Only limited photographic rules are implemented in these work and the resulting images either contain artifacts or are not very compelling. More recent state-of-the-art automatic recomposition methods (Nishiyama et al. 2009, Bhattacharya et al. 2010, Liu at al. 2010, Liu et al. 2010) have achieved more success in improving aesthetics of images. These work employed one of the three image operators; cropping, warping, or patch-relocation aka cut-and-paste to recompose an image. However, almost all these methods work well only when the subjects are already visually dominant with respect to their immediate background. None has attempted to make the subjects more dominant by directly changing the subject-background spatial relationship.

1.1 Thesis Objectives

Motivated by the important role played by visual dominance in influencing image aesthetics, and the lack of research work to automatically enhance visual dominance as a mean to improve image aesthetics, in this dissertation we focus on using a saliency-based approach for image aesthetics evaluation and enhancement. We aim to improve photographic aesthetics by modifying both the *low-level features* and *spatial composition* of an image to enhance the visual dominance of the photo subject. To ensure an image enhancement algorithm effectively increases image aesthetics and not otherwise, it is mandatory to implement an aesthetics measure to guide the image enhancement operation. For this purpose, we develop *aesthetics evaluation models* to automatically measure image aesthetics of a given photograph. The objectives of this dissertation are thus threefold:

1) Develop **saliency-based aesthetics evaluation** models for aesthetics class and score prediction.
2) Develop a **saliency-based, aesthetics-driven low-level image enhancement** method to retarget the saliency of photo subjects to coincide with the target saliency intended by users and to enhance image aesthetics.
3) Develop a **saliency-based, aesthetics-driven image recomposition** method to semi-automatically modify the spatial composition of an image to enhance visual dominance of the photo subject and other aesthetics elements.

1.2 Thesis Contributions and Their Significance

Very broadly, our contributions in this dissertation can be summarized as adopting the saliency-based approach towards aesthetics evaluation and enhancement of an image. In line with the three objectives outlined in the previous sub-section, we now present the details of the specific contributions made in this dissertation.

1.2.1 Saliency-enhanced Aesthetics Evaluation

Computational image aesthetics evaluation can be very useful in various photographic applications, such as digital photo-editing, content-based image retrieval, content-based document design, and even during photo-taking. Existing work (Tong et al. 2002, Yan et al. 2006, Datta et al. 2006) based on the computation of aesthetics features and photographic rules have shown promising results but have reached performance bottleneck, with all methods yielding about the same classification accuracy of about 70% to 72%. One underlying limitation may be that these methods focus mainly on global image features. Studies have shown that there exists strong correlation between visual attention and visual aesthetics. According to Lind (1980), aesthetic objects are interesting and thus, can hold and attract attention. Similarly, Coe (1992) discovered that aesthetics is a means to create attention to an object or a person. These studies suggest that visual attention may be a key to aid the evaluation of photographic aesthetics and improve accuracy of aesthetics model. In this work, we explore the use of higher-level perceptual information, based on visual attention, for aesthetics class and score prediction. In addition to a set of discriminative global image features, we extract a set of salient

features that characterize the subject and depict the subject-background relationship to train the aesthetics models. This high-level perceptual approach produces a promising 5-CV classification accuracy of 78.8%, significantly higher than existing approaches that concentrate mainly on global features. For the aesthetics score prediction model, despite moderate accuracy, it still shows improvement compared to existing models and is proven useful to drive the low-level image enhancement in our saliency retargeting approach presented in the section 1.2.2.

Significance:

- The idea of using a set of subject and subject-background features in training aesthetics evaluation models.
- A saliency-based, aesthetics class prediction model to discriminate between professional photographs and snapshots captured by amateurs.
- A saliency-based, aesthetics score prediction model to predict the aesthetics score of a given photograph.

1.2.2 Saliency-based Low-level Image Enhancement

In the study of photography and aesthetics, Wollen (1978) revealed that photographers deliberately avoid uniform sharpness of focus and illumination as an approach to achieve higher image aesthetics. This approach is based on the basis that our eyes are attracted to salient elements that are acutely sharp, bright or colorful in images. Figure 1.2 shows examples of how professional photographers utilize contrast in sharpness, lighting, and color to bring out the visual dominance of subjects so that the viewer is directed to where the photographers intended.

Figure 1.2. Visual dominance of the photo subject can be achieved using (a) acutely sharp focus, (b) lighting contrast, and (c) color contrast. Images courtesy of Roie Galitz (Berkeley Segmentation Dataset).

In this dissertation, we introduce a new approach to enhance image aesthetics through **saliency retargeting**. The key idea of saliency retargeting is to alter three low-level image features; intensity, color and sharpness of the objects in the photograph, such that their computed saliency measurements in the modified image become consistent with the user-intended order of their visual importance. This method generates many such modified images that satisfy the specified order of importance, and uses an aesthetics score prediction model to pick the one with the best aesthetics. The goal is to produce a maximally-aesthetic version of the input image that can redirect the viewers' attention to the most important objects in the image, and thus making these objects the main subjects. This is useful for enhancing photographs that do not have any obvious main subjects, or for photographs that one wishes to swap the role of the main subject with some other objects. Figure 1.3 shows a simple result from our method. In the original image, the intended subject (the fish) does not stand out due to the distracting background. In the resulting image, the saliency of the background has been suppressed, making it less distracting, and the fish has become more salient, making it the most dominant

subject. This shift of saliency to the intended subject is evident in the resulting saliency map. User studies performed illustrate the effectiveness of our approach in retargeting image saliency and making the retargeted image more aesthetically pleasing.

Significance:

- The idea of saliency retargeting – altering the saliency of the object(s) in a photograph to match the intended order of importance given by users.
- A simple, practical algorithm to perform saliency retargeting to alter three low-level image features; intensity, color and sharpness of the photo subjects, leading to enhanced visual dominance and improved image aesthetics.

1.2.3 Saliency-based Image Recomposition and Image Retargeting

None of the state-of-art recomposition methods (Barnajee et al. 2007, Kao et al. 2008,

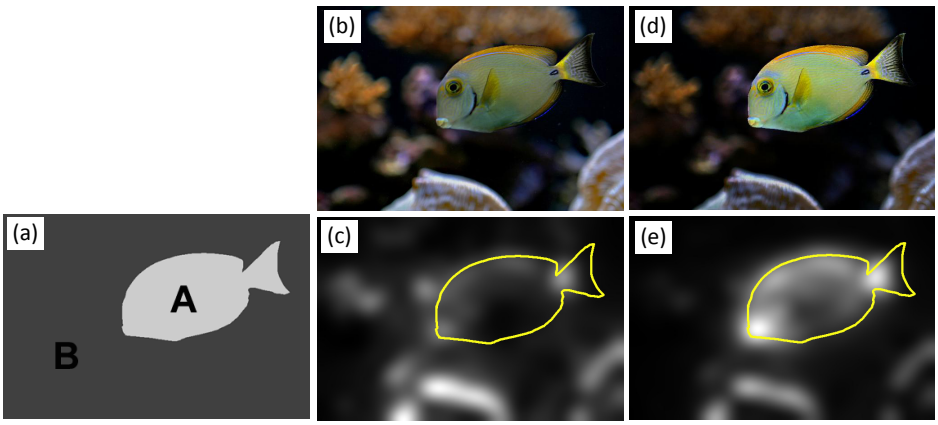


Figure 1.3: (a) Object segments, where Objects A and B are in decreasing order of importance. (b)-(c) Original image and its saliency map. (d)-(e) Image enhanced by saliency retargeting and its saliency map.

Nishiyama et al. 2009, Bhattacharya et al. 2010, Liu at al. 2010, Liu et al. 2010) aim to enhance visual dominance of the photo subjects, partly due to the unavailability of an image geometric trasnformation operator that has the flexibility to modify the spatial relationship between the subject and the background without violating spatial semantics. A significant contribution of this dissertation is a new image warping method, termed as **tearable image warping**, that can support scene-consistent image recomposition and image retargeting. In tearable image warping, we divide each selected object's boundary into tearable and non-tearable segments. Normally, the tearable segments correspond to where depth discontinuity occurs, and non-tearable segments to parts of the object boundary that have actual physical contacts with the environment or other objects. Conceptually, during warping, we allow the object's boundary to tear along the tearable segments. This allows the background to partially break away from the object and be warped more independently, which often can distribute warping more evenly to avoid local distortion. Meanwhile, the object is kept undistorted and the non-tearable segments help to preserve image semantics by constraining the object to maintain the real contacts in the 3D world. Any hole left behind after the warping is automatically inpainted (Criminisi et al. 2004, Yousef et al. 2011). The target application is image recomposition and image retargeting.

Recomposition results of tearable image warping in Figure 1.4 demonstrate the effectiveness of this approach to enhance visual dominance through the change of spatial composition between the subject and its background, while preserving the semantic connectedness of the image. In addition to making the subject dominant, other photographic rules such as rule-of-thirds, visual balance and aesthetically pleasing sizes have also been applied to improve image aesthetics. Results of our

Figure 1.4. (column 1 and 2) Input images and their corresponding saliency maps (column 3 and 4) Results of terable image warping and their corresponding saliency maps, illustrating its effectiveness in enhancing visual dominance of photo subject(s).

empirical user studies prove the effectiveness of this recomposition approach in enhancing both visual dominance and aesthetics of images. In terms of image retargeting, results show that the proposed tearable warping algorithm in general produces less distortion than the traditional non-homogeneous warping methods (Jin et al. 2010) and can better preserve scene consistency by maintaining the desired connectedness between objects and background compared to scene carving (Mansfield et al. 2010).

Significance:

- The concept of tearable/non-tearable object boundary, which leads to more flexible warping without sacrificing image semantics preservation.
- A practical algorithm to implement our tearable image warping idea for image recomposition and retargeting.

- A novel, aesthetics-driven recomposition method that capacitate the modification of the spatial relationship between subjects and the background to enhance visual dominance of photo subjects and other aesthetics elements.
- A novel retargeting method that can preserve all three scene consistency properties — object protection, correct depth order, and semantic connectedness — simultaneously in extreme retargeting cases

1.3 Thesis Organization

To provide adequate background for this thesis, Chapter 2 provides a comprehensive study on photographic aesthetics and visual saliency, two fundamental theories underpinning the research on saliency-based image enhancement. We then provide a detailed review on existing computational aesthetics evaluation models and image enhancement methods. In the subsequent chapters, we present our research work on saliency-based aesthetics evaluation and image enhancement. In Chapter 3, we present the saliency-enhanced approach for aesthetics class and score prediction. Next, in Chapter 4, we introduce saliency retargeting, a novel low-level image enhancement approach aimed to enhance image aesthetics by redirecting viewers' attention to the important subjects of the scene. In Chapter 5, we commence by depicting the algorithmic details of tearable image warping, an innovative variant of image warping that holds several advantages over pure image warping. We then present the application of tearable

image warping for scene consistent image retargeting and image recomposition. We end this chapter with the empirical evaluation to study the effectiveness of our recomposition approach. Finally, we conclude this thesis with a summary of the research work presented in this thesis and outline the future research direction inspired by this thesis.

Chapter 2

Background

Photography is more than a medium for factual communication of ideas. It is a creative art.

Ansel Adams

In this chapter, we provide a comprehensive background studies on photographic aesthetics and visual saliency, two fundamental theories underpinning the research on saliency-based image enhancement. We then explore the state-of-art computational aesthetics evaluation models and image editing methods. For image editing methods, we performed a comprehensive study into the existing work of three broad categories of image editing; low-level image enhancement, image recomposition, and image retargeting.

2.1 Photographic Aesthetics:

2.1.1 Theory and Computational Methods

The goal of this dissertation is to enhance photographs to make them more aesthetically pleasing. It is therefore important to perform a thorough study on photographic aesthetics and to establish a computational aesthetics model to guide the image enhancement process.

2.1.2 Photographic Rules and Their Aesthetics Appeal

After a comprehensive study on photographic aesthetics, we conclude that the important elements of photographic aesthetics can be grouped into six categories namely subject dominance, emotion, light and color, focusing control, balance and geometric elements, as illustrated in Figure 2.1. Among these aesthetics elements, subject dominance is arguably the most important component and is therefore placed in the centre of the diagram. A photograph with a visually dominant subject in general induces stronger aesthetic interest. Vice versa, a photograph without a dominant subject or one with more than one dominant center of interest can be puzzling to a viewer, leading to decreased aesthetics experience.

Professional photographers employ a rich set of **photographic rules** to enhance at least one of these aesthetics elements to make their photographs more appealing. These photographic rules may involve changing the *composition, exposure or depth of field* of a snapshot by adjusting the *camera position / orientation / view angle, zoom, shutter speed, or aperture*. It is important to note that each photographic rule may carry different weight for different type of photographs. For instance, low depth of

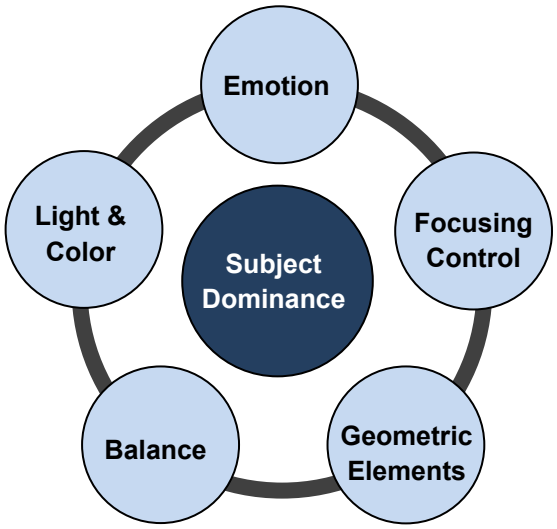


Figure 2.1. Elements of photographic aesthetics.

field is desirable for portrait but not for landscape where we want all elements sharp. Vice versa, framing and rule of thirds is not so significant to portrait and macro photography since the subject may fill up the frame for a close up. In the following sub-sections, we provide the detail description of a set of photographic rules categorized by the aesthetics element that it aims to enhance.

2.1.2.1 Subject Dominance

It is pertinent to make subject(s) of interest dominant so that viewers' attention is directed to what a photographer wants them to see. This explains why many of the photographic composition rules are targeted to increase the dominance of the main subject(s).

Simplicity: *Simplicity* is an utmost important rule that professional photographers are faithful to. Professional photographers achieve simplicity by choosing a camera view angle such that the background behind the photo subject is simple, making the

photo subject more dominant. In Figure 2.2(a), we can observe that the good choice of camera viewpoint chosen by a professional photographer makes the photo subject distinctively more visually dominant. Comparatively, the giraffe in the snapshot captured by a casual photographer in Figure 2.2(b) does not stand out due to the distracting background.

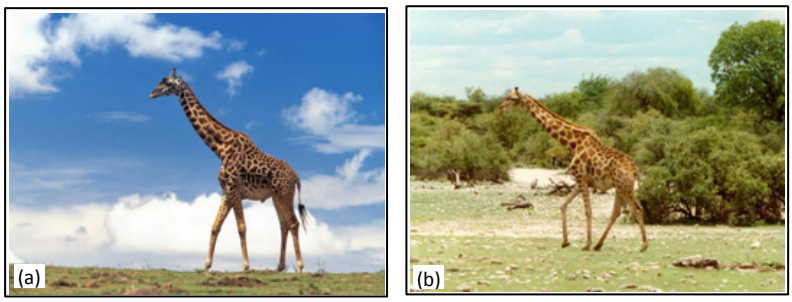


Figure 2.2. Photographs captured by (a) a professional photographer, and (b) a casual photographer.

Fill the frame, framing and leading lines: *Filling the image frame* with the photo subject eliminates distraction surrounding the subject and allowing viewers to focus fully on the photo subject. *Framing* and *leading lines* are two popular artistic techniques used by photographers to direct viewers' attention to the photo subject. Some examples that follow each of these rules are illustrated in Figure 1.1.

2.1.2.2 Equilibrium – Our need for Balance

The principle of *equilibrium* explains our search for balance in everything we see. Our visual judgments are greatly influenced by balance. A balanced picture is deemed to be more aesthetically pleasing to the eyes. There are two types of balance, **symmetric balance** and **asymmetric balance**. Reflection of the landscape in still water is an example of almost perfect symmetry. However, in most situations,

asymmetric balance, sometimes called dynamic balance is considered more pleasing in a photograph than symmetric balance. In photography, balance can be achieved using rule-of-thirds and visual balance.

Rule of Thirds (Golden Ratio): The rule of thirds, a photographic composition rule based on the approximation of the golden ratio used in artistic paintings, is used to place the elements of interest such that it indirectly contributes towards an asymmetrically balanced image. It works amazingly in drawing the human attention into the composition. According to this rule, objects should be placed near one of the four power points, which are the intersections of the two vertical and two horizontal lines that divide the image into nine equal rectangular regions. In addition to the object's location, the rule of thirds is often applied to positioning the horizon, where it is placed near one of the two horizontal power lines. Figure 2.3 illustrates examples of photograph adhering to this rule.

Figure 2.3. (left) Photo subject is placed near the bottom right power point (yellow). (right) The horizon is placed near the top power line (red).

Visual Balance: Visual balance builds upon the notion of visual weight, where an object is visually heavier if it is larger and more salient. In other words, placing the main subject off-center and balancing the "weight" with other objects. In a visually balanced image, the center of the "visual mass" is close to the center of the image. Balance can be achieved symmetrically or asymetrically as illustrated in Figure 2.4.

Figure 2.4. Photographs illustrating (left) symmetrical balance (Photo by Fabio Montalto) and (right) asymmetrical balance.

2.1.2.3 Geometrical Elements

Characteristics of geometrical elements also influence the aesthetic judgment of a photograph.

Image perspective: In photography, three-dimensional world is being rendered onto a two-dimensional image. Therefore, image perspective is very important as it can reproduce a strong sense of depth. One artistic way to show perspective is taking a photograph with converging parallel lines. For example, the parallel lines of a railway track in Figure 2.5 are perceived to converge at a distant vanishing point in the horizon.

Lines, curves and shapes: *Diagonal lines*, including *leading lines* have strong aesthetic appeal. *S-Curve* is another compelling compositional element. It adds a

Figure 2.5. Photographs following different rules to include aesthetically pleasing geometrical elements; (a) diagonal line, (b) perspective, (c) perspective / leading line, and (d)-(e) S-curve.

sense of movement to an otherwise static image. S-curves can be created by objects such as stream, path, railing, and curved object as illustrated in examples in Figure 2.5.

2.1.2.4 Light and Color

Exposure of Light: Apart from special cases where over-exposing or under-exposing a photograph can lead to a specific desired effect, we seek to capture a photograph with "correct" exposure. However, obtaining the "correct" exposure can be very tricky and subjective at times because the real world contains a wider range of tones than even the best digital sensors can represent. Good contrast is another important feature in determining the aesthetics value of a photograph. Examples of pictures with good exposure and contrast are shown in Figure 2.6.

Figure 2.6. Photographs with (left) good exposure (Photo courtesy of Philip Greenspun) (right) good contrast (Photo courtesy of Ansel Adams)**.**

Color: Except when working in a studio, photographers can seldom choose their color palette. However, photographers can sometimes change their viewpoint to obtain desirable color combination. There are two preferred color combination: **harmony of similarity** and **complementary harmony** (Freeman, 2007). Harmony of similarity describe that analogous colors, colors adjacent to each other in the color palette, produce a soothing effect when put together as illustrated in Figure 2.7a.

On the other hand, Figure 2.7b shows that complementary colors, colors directly opposite to each other, have the ability to enhance the contrast of an image.

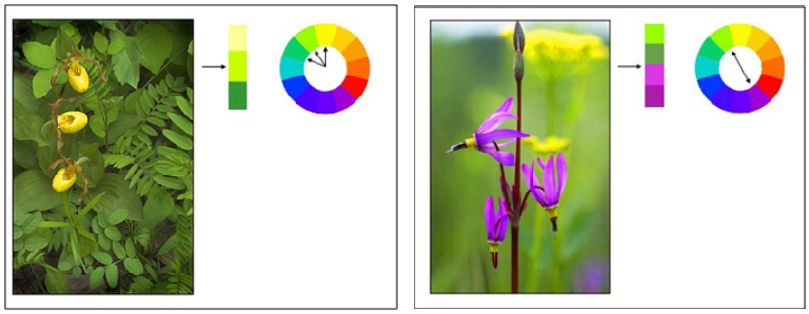


Figure 2.7. Color harmony based on YRB color palettes (left) Harmony of similarity (right) Complementary harmony (Berdan, 2004).

2.1.2.5 Focusing Control

Focusing control determines the depth-of-field, the area in a photograph where the objects are sharp and on focus. In some photos such as macro and portrait, **low depth-of-field** is desirable to place more emphasis on the photo subject, making it the only object in focus. On the other hand, images such as landscape require **high depth of field** to provide front-to-back sharpness. Figure 2.8 shows two professional photographs with low and high depth-of-field.

Figure 2.8. Focusing control. (left) Macro – low depth-of-field; only the bird is sharp and in focus (right) Landscape – high depth-of-field; whole image is sharp.

2.1.2.6 Emotion

Emotion, or feeling, is another important ingredient that makes a photograph shines. A great photo stimulates viewers' emotional response and connects viewers with the photograph. Emotion can be portrayed through the face expression of the photo subject. For example, the left image in Figure 2.9 successfully captures the spontaneous sense of awe and joy of an innocent child. Alternatively, a photograph can also convey feelings such as melancholy, gloom, sadness, and desolation through the emotional environment of a captured scene. The image of a foggy, deserted city in Figure 2.9 undoubtedly invokes a sense of desolation and loneliness. However, we exclude the study of emotion in our work as it encompasses high level of subjectivity and semantic analysis that does not fit into the scope of this thesis.

Figure 2.9. Emotion in photographs. (left) An innocent boy in a joyous mood. (right) The foggy deserted city portrayed a sense of desolation and loneliness.

2.1.3 Approaches for Evaluating Visual Aesthetics

To ensure the aesthetics of the image is improved after performing saliency-based image enhancement, we propose maximizing aesthetics of the output image as one of the objective function of the optimization problem. Therefore, there is a need for

an approach that can compare and evaluate the visual aesthetics of two images. Research on visual aesthetics evaluation is a pretty new field of research with only a handful published work. Most of the existing work focuses on classification of photographs to either good or bad photographs. Research work on score prediction is rare.

Classification of photographs based on aesthetics measures was first attempted by Tong et al. (2002), in which they took a black-box approach to classify photographs into *professional* or *snapshots*. A large set of 846 low-level features were combined exhaustively with a standard set of learning algorithms for classification. Although this approach successfully classifies photographs with an accuracy significantly better than chance, it offers little insight into why certain features are selected, or how to design better features. Yan et al. (2006) tried to address the above limitations by using a principled approach. They studied the perceptual criteria that people use to judge a photo and presented a top-down approach to construct high-level semantic features for assessing the quality of the photos. With a small set of highly discriminative high-level semantic features, they achieved a classification accuracy of 72.3% using a Naïve Bayes classifier, an accuracy comparable to that of Tong et al.'s approach.

In a similar work, Datta et al. (2006) computed a set of 56 features based on rules of thumb in photography, common intuition and observed trends in ratings. Combining filter-based and wrapper-based methods, they shortlisted a set of 15 features and used them to classify photos into '*high*' and '*low*' classes. Using a set of photos from Photo.net, with aesthetics scores ranging from one to seven, and excluding photos with average scores between 4.2 and 5.8, they obtained a classification accuracy of 70.12% using an SVM classifier.

Comparing these approaches, Yan's and Datta's approaches are more objective and efficient for aesthetics class prediction. Yan's has a smaller set of more discriminative features compared to Datta's larger but weaker set of features. However, all methods obtained about the same classification accuracies even though different sets of features and classifiers have been used. A possible bottleneck of these approaches is that none of these methods consider features specific to the photo subject, which potentially provides insight into a better set of discriminative features.

The attempt to predict aesthetics score was pioneered by Datta el al. (2006). Using the same features and dataset they used for class prediction, they performed linear regression on polynomial terms of the feature values to predict the aesthetics score for images. They only manage to achieve a residual sum-of-squares error of 0.502 which is a reduction of only 28% from the variance. Although the score is not good enough for practical use, it suggests that visual features are able to predict human-rated scores with some success. More recently, Kao et al. (2008) derives a method to compute composition score based on a set of five selected photographic rules. This set of rules does not cover many aesthetics elements especially the low-level features of photo subjects such as contrast, saturation and texture variation. In summary, research on predicting the aesthetics score of an image is still in its infancy stage and warrants investigation.

2.2 Visual Saliency: An Important Element of Photographic Aesthetics

In the previous section, we have identified visual dominance of a photo subject as the most important aesthetics element in photography. In order to enhance visual

dominance in an image, there must be a way to measure visual dominance as a relevance feedback to the image enhancement algorithm.

2.2.1 Approaches for Determining Visual Saliency

To perform saliency-based image enhancement, a method that can determine the contrast of image regions to their surroundings is needed. Existing saliency estimation methods can be classified as biologically-inspired, purely combination, or a combination. These methods use one or more features of intensity, color, and orientation to determine the saliency of an image.

We look into five state-of-the-art methods selected based on citation in literatures, recency, and variety; Itti et al (1998), Ma and Zhang (2003), Hou and Zhang (2007), Harel et al. (2007) and Achanta et al. (2009), referred to as IT, MZ, SR, GB, and IG respectively. Table 2.1 shows the comparison among these methods. IT is the classical algorithm that is built upon a biologically plausible architecture (Koch and Ullman, 1985). Multi-scale features are combined into a single topographical saliency map through the activation and normalization steps. The activation step is accomplished by subtracting the respective feature maps of different scale. The normalization is performed based on a local maxima scheme, which promotes maps where a small number of strong peaks are present and suppress maps that contain numerous comparable peak responses to obtain the final saliency map. Despite being the oldest method, Itti's biologically inspired model remains the most popular method and has been used in a number of image enhancement and re-composition methods (Setlur et al. 2007, Wang et al. 2007, Kao at al. 2008). GB is based on the same biological model as IT and uses the same set of initial features maps but it replaces the activation and normalization step with a

graph-based approach. They defined Markov chains over the features maps and treat the equilibrium distribution over map locations as activation and saliency values. Based on their experimental results on 749 variations of 108 natural images, GB predicts human fixation with higher accuracy, achieving 98% of the ROC area over a human-based control, compared to IT method that achieves only 84%.

MZ, SR, and IG are purely computational methods. MZ proposed a single-scale method based on local contrast analysis. The input to their algorithm is a resized and color-quantized CIELuv image. The saliency map is obtained by summing the differences of the image pixels with the respective surrounding pixels in a small neighborhood. To simulate the human visual perception, a fuzzy growing method is used to compute the attended areas. SR method is a simple and fast method for saliency detection that is independent of features, categories, and other form of prior knowledge about the image. The input image is resized to 64 x 64 pixels. By analyzing the log spectrum of an image, they extract the spectral residual in the spectral domain and proposed a fast method to construct the saliency map in the spatial domain. Finally, IG introduces a frequency-tuned method to estimate center-surround contrast using only color and luminance features. The advantages of IG over the other methods are it produces uniformly highlighted salient regions with well-defined boundaries, full resolution saliency maps, and is computational efficient.

Table 2.1 depicts the comparison of the saliency estimation methods. Among these methods, only IT and GB are biologically inspired. IT, GB, and MZ consider all three low-level features of intensity, color and orientation whereas IG only exploits the color and luminance features and SR is independent of features. IT, GB and MZ produces intermediate individual saliency map for each feature but are less

computationally efficient compared to SR and IG that do not generate any individual feature map. Figure 2.10 depicts the quantified performance comparison with the precision-recall curve for naïve thresholding of saliency maps on 1000 images. This performance comparison only compares how well the saliency map covers the full subject and not how accurate it is in predicting human fixations. On this ground of comparison, IG gives the highest precision followed by GB and this result is evident from the visual comparison illustrated in Figure 2.11. A notable observation is that IT achieves very high precision with low recall but the performance drops steeply with increase in recall. This is explainable from the saliency map that is generates because IT only detects parts of an object that attract attentions but does not covers the full object as illustrated in the visual comparison of the methods in Figure 2.11(b).

In summary, despite being computationally efficient and perform better than other methods, IG does not provide sufficient clue for our image enhancement task

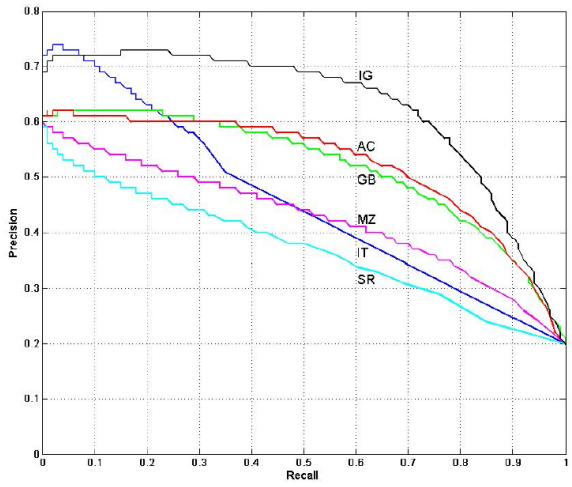


Figure 2.10. Comparison of approaches for determining visual saliency. Precision-recall curve for naïve thresholding of saliency maps on 1000 images (Achanta, 2009).

Table 2.1: Comparison of saliency estimation methods

Saliency estimation method	Brief description	Type	Low-level features used			Produce individual saliency maps?	Resolution of saliency map (S = image size)	Efficient??
			Intensity/ Luminance	Color	Orientation			
IT (Itti et al, 1998)	Difference of Gaussian (DoG)	Biologically-inspired	√	√	√	Yes	S/256	Moderate
MZ (Ma & Zhang, 2003)	Center surround feature distance with fuzzy growing	Purely computational	√	√	√	Yes	S/100	No
SR (Hou & Zhang, 2007)	Spectral-residual approach	Purely computational	Independent of features			No	64 x 64 pixels	Yes
GB (Harel et al., 2007)	Graph-based propagation	Combined	√	√	√	Yes	S/64	No
IG (Achanta et al., 2009)	Frequency-tuned approach	Purely computational	√	√	-	No	S	Yes

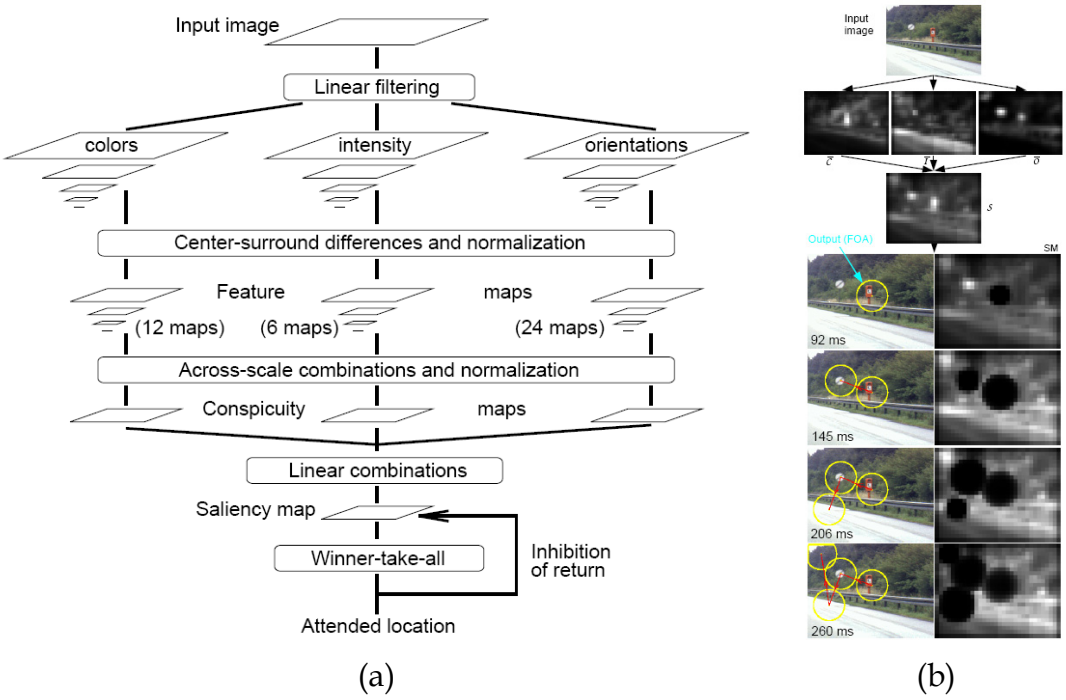
Figure 2.11. Visual comparison of saliency maps (Achanta, 2009).

because it uses only two features and does not produce any individual saliency map for its features. The more promising method for this context would be GB and IT that generate intermediate feature map for all three features and gives comparable performance both visually and quantitatively. Comparatively, GB has higher accuracy compared IT but lower efficiency in generating the saliency maps. Another advantage of IT is that it can produce a sequence of predicted eye fixation locations which can be very useful. Since IT and GB are based on the same biological plausible architecture, we will use either IT or GB in this dissertation depending on the application context.

2.2.2 Itti-Koch Visual Saliency Model

In this section, we detailed the biological architecture (Itti and Koch, 1998) underlying IT and GB models in order to provide insights to the adaption of this model in the saliency-based image enhancement methods proposed in Section 4 and 5. Itti-Koch developed a biologically plausible computational model of visual attention with emphasis on bottom up control of attentional deployment. This model tries to reproduce the behavior of human visual system. Features used in this model are color, intensity and orientation.

In visual attention-based approach, the main content is technically termed as the visual attention region (VAR). The general architecture of the visual saliency model is shown in Figure 2.12(a). First, the feature maps are extracted from the image. Altogether, 42 features are computed to build the saliency map; 12 color maps, 6 intensity maps and 24 orientation maps. Each feature is computed by a set of linear "center-surround" operations, implemented as the difference between fine and



(a) (b)

Figure 2.12. (a) General architecture of the visual saliency model (b) Example of operation of the model with a natural image (Itti et al. 1998).

coarse scales created using dyadic Gaussian pyramids (Greenspan et al. 1994). The center is a pixel at scale $c \in \{2, 3, 4\}$ and the surround is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{3, 4\}$. The center surround differences (Θ) between a "center" of a fine scale and the "surround" of a coarse scale yield the feature maps. The first set of 6 intensity maps for each center-surround combination is given by:

$$I(c,s) = | I(c) \Theta I(s) | \qquad (2.1)$$

A second set of 12 feature maps is represented using the "color double-opponent" system. Maps RG(c,s) are created to account for red/green and green/red double opponency and BY(c,s) are created for blue/yellow and yellow/blue double opponency:

$$RG(c,s) = |\ (R(c) - G(c))\ \Theta\ (G(s) - R(s))\ | \tag{2.2}$$

$$BY(c,s) = |\ (B(c) - Y(c))\ \Theta\ (Y(s) - B(s))\ | \tag{2.3}$$

Lastly, the local orientation information is obtained from I using oriented Gabor pyramids $O(\sigma, \theta)$, where $\sigma \in [0..8]$ represents the scale, and $\theta \in \{0°, 45°, 90°, 135°\}$. The local orientation contrasts between the center and surround scales are represented by 24 orientation maps, $O(c, s, \theta)$:

$$O(c,s,\theta) = |\ O(c,\theta)\ \Theta\ O(s,\theta)\ | \tag{2.4}$$

These feature maps are normalized by a map normalization operator, N and then combined into three *conspicuity maps* (Figure 2.12(b)), $\bar{I}$ for intensity, $\acute{C}$ for color and $\bar{O}$ for orientation. They are obtained through across-scale addition, "$\oplus$" by reduction of each map to scale four and point-by-point addition. The *conspicuity maps* are given by:

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N\ (I(c,s)) \tag{2.5}$$

$$\acute{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N\ (RG\ (c,s)) + N\ (BY\ (c,s))] \tag{2.5}$$

$$\bar{O} = \sum_{\theta \in \{0°,45°,90°,135°\}} N\ (\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N\ (O(c,s,\theta))) \tag{2.6}$$

The three conspicuity maps are normalized and summed into S the final input to the *saliency map (SM)*:

$$S = \frac{1}{3}(N(\bar{I}) + N(\acute{C}) + N(\bar{O})) \tag{2.7}$$

The maximum value of saliency map (SM) defines focus of attention (FOA) location. The SM is feed into a biologically plausible 2D "winner-take-all" (WTA) neural

network in which only the most active location remains, while all other locations are suppressed. At any given time, the maximum value of SM defines focus of attention (FOA) location as shown in Figure 2.11b.

2.3 Computational Methods for Image Editing

2.3.1 Low-level Image Enhancement

Traditonal automatic image enhancement, including contrast enhancement (Tomasi and Manduchi, 1998, Rahman, 2004), color correction and balancing (Barnard et al., 2002, Moroney, 2000, Rizzi et al., 2003), and edge sharpening (Kashyap, 1994, Polesel et al., 2000) alter only the global features or local features based on neighborhood information, without considering the content of a photograph, partly due to the non-trivial recognition of the photo subject. In addition, these methods are targeted to deal with image degradations rather than to enhance image aesthetics.

More recently, several aesthetics-driven image enhancement techniques for manipulation of image tone and color and sharpness have emerged. Bae et al. (2006) introduced a tone management approach that allow users to dictate the 'look' of their images by transferring distinctive toning styles from professional photographs, such as those captured by master black-and-white photographers. Cohen-Or et al. (2006) presented a color harmonization method that enhances the harmony among the colors of a given photograph by shifting the colors towards a harmonic setting.

Su et al. (2005) first proposed the idea of altering saliency of an image by reducing the background saliency to redirect attention to the main subject. Their method utilizes texture power maps to de-emphasize texture variations to decrease

the salience of distracting regions. This method preserves key features, but while adding white noise maintains overall graininess, the resulting images appear too noisy. Gasparini et al. (2007) performed selective edge enhancement by implementing the unsharp mask weighted by the saliency map of the image. Both of these work managed to make the photo subject more dominant but the resulting image is not necessarily aesthetically more pleasing because no appropriate constraint or photographic rules is applied to ensure the aesthetics aspect of the image is enhanced. Barnajee et al. (2007) and Bae and Durand (2007) achieved more success in enhancing both subject dominance and image aesthetics by simulating the shallow depth of field effect, an aesthetics feature often desired for photographs such as portraits. Barnajee's approach detects the photo subject using the out-of-focus information from a supplementary image and given the location of the subject, blurs the entire background to simulate the shallow depth of field effect. However, practicality of this approach is limited as it requires an additional supplementary image with sufficient defocus information in order to extract the subject. Bae and Durand used a more robust, defocus magnification approach. Their approach estimates the spatially-varying amount of blur over the image and then magnifies the existing blurriness by blurring the blurry regions and keeping the sharp regions sharp. This approach successfully simulates the low depth of field effect and produces reasonably good results although some artifacts do exist such as halo effects and sharp region being wrongly blurred. However, noticeably, all these approaches increase the saliency of the photo subject(s) by modifying either the texture variation or sharpness contrast of an image. None has attempted to alter intensity or color contrast between the subject and its background to enhance the visual dominance of the subject.

2.3.2 Image Recomposition

Aesthetics-based photo recomposition is an emerging area of research in the field of computational photography. Recent state-of-the-art methods all take the approach of proposing a set of photographic composition rules and employing optimizations that modify the input image so that the composition rules are adhered to as much as possible. In general, existing recomposition methods modify the input image using three spatial image operators; cropping (Barnajee et al. 2007, Kao et al. 2008, Nishiyama et al. 2009), warping (Liu et al., 2010), and patch relocation aka cut-and-paste (Bhattachary 2010).

The most traditional approach of digital image recomposition is cropping, largely due to its simplicity and artifact-free nature. Barnajee (2007) used out-of-focus blur information from a supplementary image to determine the ROIs of the image and cropped an image to conform to rule-of-thirds. In another attempt, Kao et al. (2008) utilizes a visual attention model (Itti et al, 1998) to extract the ROIs and performed automatic re-composition by using a composition scoring system to guide the cropping and rotational correction of the image. In both work, limited composition rules are implemented and results produced are not compelling. Nishiyama et al. (2009) improved on the previous aesthetically-driven cropping methods by training an aesthetics assessment model with a large collection of photos and used this model to maximize aesthetics of the resulting cropped image to produce more promising results. More recently, using a pure warping image operator, Liu et al. (2010) formulated a set of aesthetics energies based on selected photographic rules to guide the warping process to modify composition of an image to make an image more aesthetically pleasing. On the other hand, Bhattacharya et al. (2010) developed a framework for photo-quality enhancement

using the patch relocation operator, where foreground objects are extracted and then pasted onto new background locations. They trained their aesthetic measurement score using real user data to guide the semi-automatic image recomposition.

Comparing these image recompositon operators, cropping and pure warping can preserve image semantics reasonably well because the topology of the spatial arrangement of objects is preserved. However, the ability to rigidly preserve image semantics limits their flexibility for more significant recomposition. For example, warping and cropping alone cannot remove unpleasant object merger (such as a tree branch sticking out behind a person's head). On the other hand, patch relocation is more flexible to deal with some cases that warping would have failed. However, without suitable constraints, patch relocation can easily produce semantically illogical results. Existing recomposition methods based on patch relocation (Bhattacharya et al. 2010) are still too simplistic or lack an adequate set of constraints, and can only be applicable to limited types of images.

Due to inherent limitation of approaches using single image operator, recent work attempt to employ hybrid operators to recompose an image. Based on well-grounded photo composition rules, Liu et al. (2010) proposed a computational means to measure composition aesthetics and utilized a crop-retarget operator that combines cropping and warping to recompose an image. Particle swarm optimization method is used to produce a maximally-aesthetics version of the input image. This approach capitalizes on the strength of both cropping and pure warping to successfully produce promising results for a wider range of images. However, both cropping and warping operators inhibit change of the background surrounding a photo subject, limiting its support for significant recomposition such

as modifying composition to increase simplicity and enhance subject dominance.

2.3.3 Image Retargeting

There are three popular classes of content-aware image retargeting operators: cropping, seam-carving, and warping. Each of these approaches utilizes some region-of-interest (ROI) extraction methods, such as saliency detection or gradient energy, to compute an importance map and tries to minimize distortion in the important regions. None of these operators can completely solve the image retargeting problem (Vaquero et al. 2010). Each operator has its own advantages and disadvantages for different applications.

Basically, content-aware cropping methods (Suh et al. 2003, Santella et al. 2009) search for the best cropping window that contains all the important objects. Some methods (Zhang et al. 2005, Nishiyama et al. 2009) try to incorporate aesthetics measures into its cropping optimization to enhance the aesthetics of the retargeted images. Cropping is artifact- and distortion-free but it may destroy the global context and is highly inflexible. Seam carving (Avidan and Shamir, 2007, Rubinstein et al. 2008) is an elegant approach that removes the least important vertical or horizontal seams, measured by a computed importance map. Although seam carving can produce some good and interesting results, it is only effective for images with sufficient homogenous regions. For more complex images, results are prone to distortions and artifacts when the seams cut through high-frequency content (Vaquero et al., 2010).

Various variants of the image warping approach have been proposed for image retargeting. The more recent warping methods represent an input image as a quad mesh (Wang et. el. 2008) or a triangle mesh (Jin et al. 2007, Guo et al. 2009) and

perform optimization to find the new locations for the mesh control points, keeping the ROIs as rigid as possible. Jin et al. (2007) went a step further to enhance image aesthetics of the retargeted images by including a set of selected photographic rules in the warping optimization. However, due to the non-changeable object-background relationship, there is limited flexibility to enhance image aesthetics using warping.

More recent approaches use a combination of multiple operators. The challenge is to find the best way to combine them. A few researchers (Dong et al, 2009, Rubinstein et al. 2009) have attempted to combine seam carving with other retargeting operators such as scaling and cropping. These methods may reduce distortions caused by pure seam carving but in cases of extreme resizing, severe distortions are largely unavoidable. Liu et al. (2010) combined cropping and warping to retarget and optimize photo composition simultaneously to produce some interesting results. However, the algorithm is slow due to its search in a 4D space and it is unable to preserve global context.

Notably, only a few retargeting methods can protect objects to avoid unpleasant distortion of objects in extreme retargeting cases. Setlur et al. (2007) and Mansfield et al. (2010) have presented retargeting methods that use a model decomposition approach to protect objects. Setlur et al. (2007) proposed a non-photorealistic retargeting method that identifies the ROIs, removes them, inpaints the holes left by the ROIs, resizes the inpainted image and re-inserts the ROIs. Relatively, this approach is highly flexible but is not popular due to its dependence on inpainting and its susceptibility to semantics violation. It also has no flexibility in repositioning the objects. Mansfield et al. (2010) proposed scene carving, a layered approach to seam carving to prevent seam from eliminating important objects. Given an image

with object segments and their respective depth order, the scene carving algorithm first decomposes the image into layers, followed by finding the best seams to remove from the background layer and positioning the objects so that their visibility is maximized. This method successfully reduces disturbing distortion to objects and can be used for extreme retargeting by allowing objects to overlap in the correct depth order. However, scene consistency may still be violated because this approach does not guarantee consistency of semantic connectedness between an object and its environment, especially when shadows or reflections exist. In addition, this approach is still prone to visual distortion inherent to the seam carving method.

2.4 Chapter Summary

In this chapter, we have outlined six aesthetics elements pertinent to photographic aesthetics; subject dominance, balance, geometrical elements, light and color, focusing control, and emotion with subject dominance being the most desired aesthetic element. The scope of this dissertation attempts to develop aesthetics evaluation models and image enhancement techniques based on all these aesthetics elements except emotion.

In general, to digitally enhance image aesthetics, aesthetically-driven image enhancement approaches attempt to mimic professional photographs by performing either **low-level image enhancement** or **image recompositon** to make an image adheres to one or more photographic rules. Table 2.2 summarizes the semi-automatic or automatic image enhancement methods based on the enhancement techniques. Notably, despite subject dominance being a much sought

after aesthetics element by professional photographers, studies on state-of-art image enhancement techniques reveal that very few research work aim to enhance the visual dominance of a photo subject.

Table 2.2. Summary of semi-automatic/automatic image enhancement methods.

Type of image enhancement		**Low-level Image Enhancement**			**Re-composition**
		Contrast	**Color**	**Sharpness**	
Global		√	√	√	
Local					
	Neighborhood info	√	√	√	
	Subject-background relationship (to enhance subject dominance)	**X**	**X**	**LIMITED** Su et al. (2005) Gasparini et al. (2007) Barnajee et al. (2007) Bae and Durand (2007)	**X**
Photographic rules except subject dominance.					**LIMITED** Barnajee et al. (2007) Wang et al. (2008) Kao et al. (2008) Nishiyama et al. (2009) Liu et al. (2010) Bhachatraya et al. (2010)

Analysis shows that most of the existing low-level image enhancement methods enhance one or more global or local features, mainly based on neighborhood information. There is only a handful image enhancement methods that consider altering the subject-background relationship. An integrated algorithm or framework that enhances all three low-level image features; contrast, color and

sharpness to make a subject dominant is non-existence. In terms of image recomposition, none of the state-f-art methods aim to improve aesthetics by making the photo subject more dominant. This could be largely due to the lack of an image operator that has the flexibility to capacitate change in subject-background relationship without violating spatial image semantics.

Motivated by the limitations of state-of-art aesthetics evaluation models and image enhancement approaches, this dissertation adopts the saliency-based approach to improve the performance of aesthetics evaluation and image enhancement. More specifically, we propose two novel saliency-based image enhancement techniques; **saliency retargeting** and **tearable image warping** to modify low-level image features and spatial image composition respectively. Interestingly, the tearable image warping method which was originally designed for image recomposition works amazingly well in image retargeting, particularly in cases of extreme retargeting. In addition, on top of a set of global image features, we train **aesthetics evaluation models** for class and score prediction using a set subject-focused and subject-background relationship features. The score prediction model was proven effective in guiding the saliency retargeting algorithm to maximize the aesthetics of the resulting images. We will present the details of the aesthetics evaluation models, saliency retargeting algorithm and tearable image warping algorithm in Chapter 3, 4, and 5 respectively.

Chapter 3

Saliency-based Aesthetics Evaluation Model

Photography is more than a medium for factual communication of ideas. It is a creative art.

Ansel Adams

The photo subject of an image may be one of the most distinguishing factors in influencing the aesthetics of a photo. However, despite the importance of the dominance of a photo subject in influencing the image aesthetics, existing work (Tong et al. 2002, Datta el al. 2006, Yan et al. 2006) on aesthetics models does not put much emphasis on exploiting subject-related features in training models for evaluating aesthetics. In this work, we adopt a subject-focused, saliency-based approach for developing both the aesthetics classification and score prediction models. More specifically, we identify the salient regions in a photo and compute a

set of features within these salient regions and a set of features based on the relationship between these salient regions and the background. Combining this set of saliency-based features with a set of important global features, we train an aesthetics evaluation model for the classification of *professional photographs* or *non-professional snapshots*. We then extend this work to train an aesthetics score prediction.

3.1 Aesthetics Class Prediction

Adopting a subject-focused approach, we first train an aesthetics class prediction model. We make the assumption that the salient regions of an image contain the subject and use the salient regions to represent the subject. The salient regions are identified using the visual attention model of Itti et al. (1998), which is built upon a biologically plausible architecture.

We collected a set of peer-rated images from Photo.net, which is an online photo-sharing community. We chose Photo.net because it has a better consensus over its ratings (Datta el al, 2006). For aesthetics classification, following Yan et al.'s approach (2006) to discern *professional photos* from *snapshots*, we extracted the top and bottom 10% of the photos and assigned them as *high-quality professional photographs* and *low-quality snapshots*.

Figure 3.1 illustrates the three steps in training an aesthetics classification model; subject extraction, feature extraction and classification. Given the training image set, for each image I, we first extract the salient mask that represent the photo subject. Each image I is converted to the HSV and LUV color spaces and the resulting two-dimensional matrices are denoted, as I_H, I_S, I_V, and I_L, I_U, I_V respectively. First, we

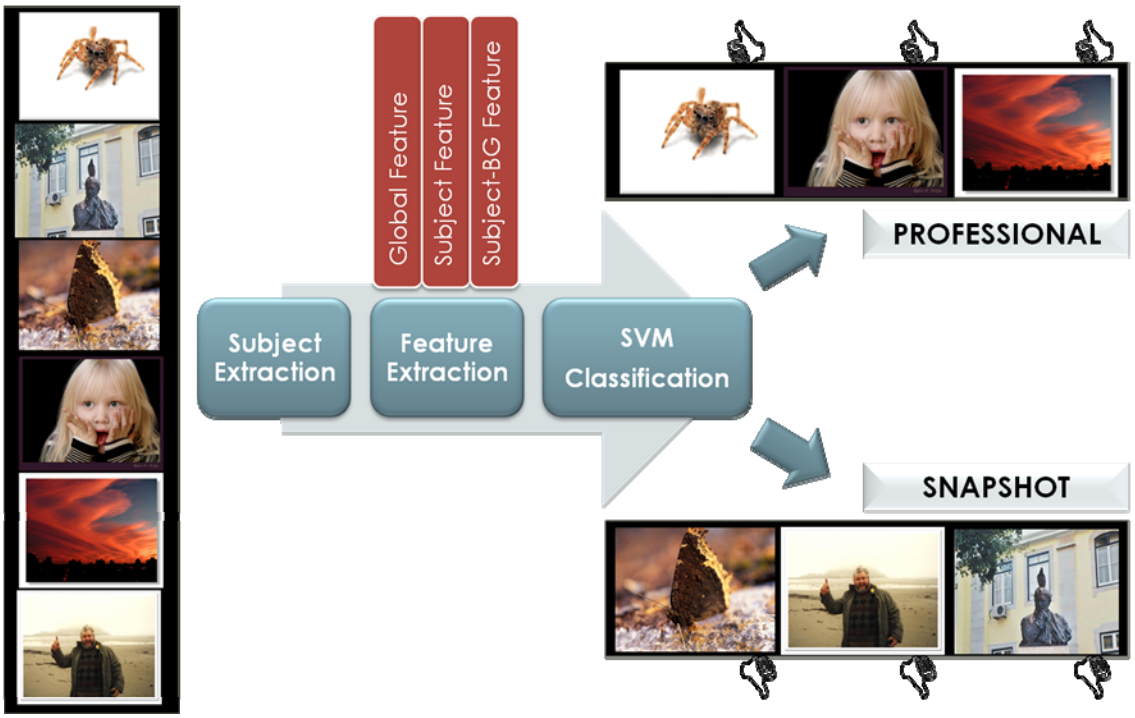


Figure 3.1. Overview of an aesthetics class prediction model.

compute the saliency map SM on I and determine a set of salient locations L. Using the salient locations in L as seeds, we perform seeded segmentation to create a salient mask K that indicates the salient regions. The salient region is defined for each HSV channel as

$$S_{ch} = \{ I_{ch}(x, y) \mid K(x, y) > 0\} \tag{3.1}$$

where ch = {H, S, V} and $I_{ch}(x, y)$ is a pixel in I_{ch}. Similarly, the background region is defined as

$$B_{ch} = \{ I_{ch}(x, y) \mid K(x, y) = 0\}. \tag{3.2}$$

Each image I and its corresponding saliency map SM, salient region S_{ch}, and background region B_{ch}, are then used to extract a set of global image features and a set of features that characterize the subject and its relationship with the background.

Altogether, we compute a total of 44 candidate features, $F = \{f_1, f_2, \ldots, f_{44}\}$. Finally, using a set of training images, each with a set of features F, we build a two-class classification model that classifies an image I into either class 1 or 0, where

$$class(I) = \begin{cases} 1 & professional\ photograph \\ 0 & snapshot \end{cases} . \qquad (3.3)$$

3.1.1 Salient Region Extraction

Extracting the main subject from a photo is non-trivial. Many methods exist but none is close to perfect or may only work well for certain type of photographs. However, some do give good hints on the salient regions of interest that attract attention and thus can be utilized to identify photo subjects.

For each image, we compute the saliency map and the salient locations using Itti's visual saliency model (1998), which is built upon a biologically plausible architecture that exploits multi-scaled intensity, color and orientation image features and learnt the salient locations using a Winner-Take-All (WTA) neural network framework. The salient locations are learnt in a sequential manner and we observed that the sequence eventually returns to the first salient location after a certain number of locations. We trace and extract only the unique salient locations. The sequence of these locations mimics the navigation pattern of the viewers, and could potentially provide clue about the image attractiveness. For our purpose, we use only the first three salient locations as seeds for segmentation as previous work (Itti et al. 1998) has shown that in over 90% of the cases, the main subject is discovered within the first three salient locations. We then perform segmentation on image, I using the CTM segmentation engine (Yang et al. 2008), a state-of-the-art segmentation technique for natural images and extract all segmented regions that

contain these three seeds to create a salient mask. This salient mask is used to create the salient region S_{ch} and its complimentary background region B_{ch} for each HSV channel. Figure 3.2 shows a photo with the salient locations and its corresponding saliency map, the segmentation result and the salient mask. The white pixels of the salient mask indicate the salient regions whereas the black pixels denote the background region.

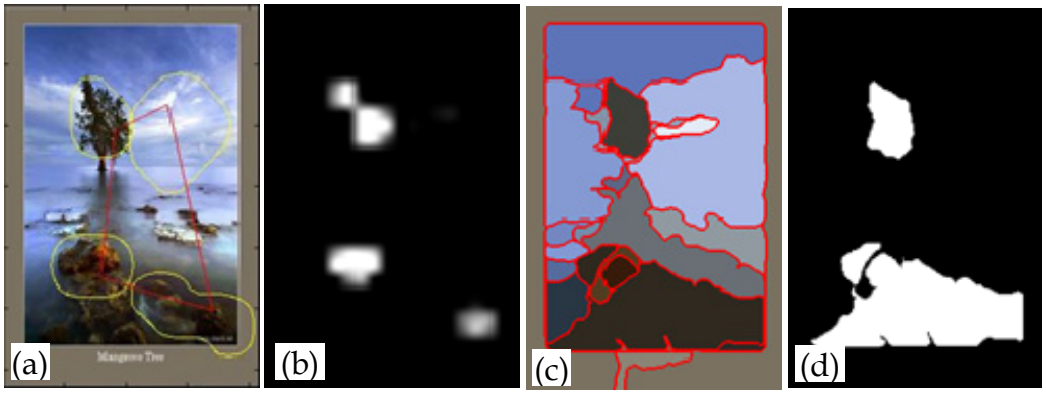


Figure 3.2. Salient image regions extraction. (a) Original image with salient locations, (b) saliency map, (c) segmented image, and (d) salient mask based on the first three salient locations.

3.1.2 Visual Features Extraction

We consider three types of features—global image features, features of salient regions, and features that depict the subject-background relationship.

3.1.2.1 Global Features

There are three categories of global image features, based on basic techniques, photographic rules, and camera settings. For basic techniques, we chose sharpness, contrast and exposure. A professional photo would be well-exposed, have a sharp

main subject, and high contrast. The blur estimation method used by Yan et al. (2006) is discriminative but rather complex with the need to combine a number of techniques: Fourier transform, Haar wavelet transform, and Naïve Bayes. We propose a simpler integrated method for detecting sharp images. Our method is based on the fact that a sharp image, including images that are partially sharp such as images with low depth-of-field (DOF), will have a significant amount of high frequencies. These high frequencies will be cut off when an image is blurred with a Gaussian filter and the total number of high frequencies being cut increases with the increase of the standard deviation of the Gaussian filter, σ. By computing the total high frequencies that are cut off when σ is increased, we can obtain the total frequencies that an image possess at different range of high frequencies and infer the sharpness of an image. First, for each image I, we create four blurred images, I_{bi} with Gaussian filters of increasing σ value,

$$I_{bi} = G_{\sigma_i} * I, \tag{3.4}$$

where

$$\sigma_1 = \min(X, Y)/(256 * \beta) \tag{3.5}$$

and

$$\sigma_{i+1} = \sigma_i \times 2 \tag{3.6}$$

for i = {2,3,4}. Then, we performed two dimensional Fourier transform on each of these four images,

$$F_i = FFT(I_{bi}), \tag{3.7}$$

and count the number of frequencies with value greater than some threshold θ,

$$C_i = \{(u,v) \,|\, \|F(u,v)\| > \theta\}. \tag{3.8}$$

We then compute the number of high frequencies, H_i, by detecting the change in the number of frequencies, C_i when σ is increased,

$$H_i = \frac{\Delta C}{\Delta \sigma} = \frac{C_{i+1} - C_i}{\Delta \sigma} \tag{3.9}$$

for i = {1,2,3). To avoid an image with a sharp subject but smooth interior and background, such as image (b) in Figure 3.3, being wrongly inferred as a blurred image, we normalize the number of high frequencies H_i with the magnitude of dilated edges, E_i. An image with dilated edges, D_i is obtained by applying the Canny edge detector on image I followed by a dilation process with a structuring element of elipse shape and size 11x11. The set of dilated edges E_i is then obtained by threshing the zeros pixels,

$$E_i = \{(x, y) \mid \|D_i(x, y)\| > 0\}. \tag{3.10}$$

Finally, the sharpness feature, f_1 to f_3 is computed as

$$f_i = \frac{\|H_i\|}{\|E_i\|}. \tag{3.11}$$

We used β = 10 and θ = 4 for our experiments. By detecting the number of high frequencies at different range, our method is able to differentiate the degree of sharpness of an image in which sharper images will have higher values for these two features. Comparatively, f_2 and f_3 is more discriminative than f_1, likely due to the existence of noise in the highest range of frequencies. The sharpness feature, f_2 calculated for images (a), (b) and (c) in Figure 3.3 are 0.512, 0.231 and 0.062 respectively. If the magnitude of high frequencies, H_i is normalized by the size of image instead of the magnitude of the dilated edges, image (b) would obtain a isharpness value of 0.054 that is obviously wrong. Our method is as discriminative as the method used by (Yan et al. 2006), if not more discriminative.

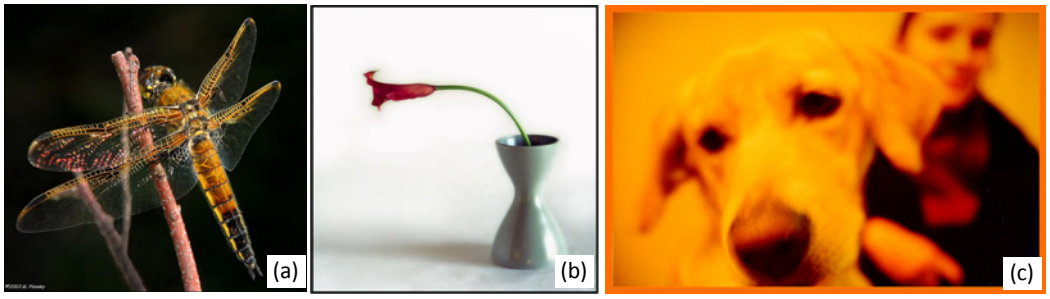


Figure 3.3. Computation of sharpness feature. (a) A sharp, low DOF image taken with macro lens; sharpness score = 0.512. (b) A sharp image with plain background and smooth interior; sharpness score = 0.231. (c) A blur image;- sharpness score = 0.062.

We modified Yan et al.'s method (2006) to compute image contrast, f_4, by taking the middle 98% mass of the luminance I_L histogram, instead of using the combined RGB histogram. The luminance records how light intensity is perceved by the human eye and therefore the I_L histogram is a more intuitive representation of image contrast. For measure of exposure, we compute the average pixel intensity, f_5 (Datta et al. 2006) as

$$f_5 = \frac{1}{|I_V|}\sum_x\sum_y I_V(x,y). \tag{3.12}$$

For the photographic rules, we measure the texture details, the low depth DOF, and the rule of thirds using Datta's method (2006). The texture details and low DOF features are computed using Daubechies wavelet transform. We performed a *three-level* wavelet transform on all three color channels; hue I_H, saturation I_S and intensity I_V. Figure 3.4(b) illustrates an example of three-level Daubechies wavelet on the intensity channel I_V. Arrangement of the four coefficients per level, *LL*, *LH*, *HL*, and *HH* is shown in Figure 3.4(c). We compute 12 features for texture details. The average of wavelet coefficients on hue channel, I_H are computed as

$$f_{5+i} = \frac{1}{S_i}\left(\sum_x\sum_y w_i^{HH}(x,y) + \sum_x\sum_y w_i^{HL}(x,y) + \sum_x\sum_y w_i^{LH}(x,y)\right) \quad (3.13)$$

where w_i^{hh}, w_i^{lh} and w_i^{hl} denote the coefficients (except LL) in level i for the wavelet transform on the hue channel I_H and i = {1, 2, 3}. Similarly, we compute the corresponding wavelet features for saturation I_S and intensity I_V channels to obtain f_9 through f_{11} and f_{12} through f_{14} respectively. Additional three texture features are formed by summing the average wavelet coefficients over all three frequency levels for hue, saturation and intensity: $f_{15} = \sum_{i=6}^{8} f_i$, $f_{16} = \sum_{i=9}^{11} f_i$ and $f_{17} = \sum_{i=12}^{14} f_i$. The low-DOF and rule-of-thirds features are computed only for the saturation I_S and intensity I_V channels. To compute the low DOF features, we divide the image into 16 rectangular blocks, {M_1, ... M_{16}}, numbered in row-major order and denote the set of wavelet coefficients in the high-frequency (level 3) as w_3 = {w_3^{hh}, w_3^{lh}, w_3^{hl} }. Assuming the subject of interest in the macro shot is usually sharp in the centre, we compute the low DOF feature for saturation as follows,

$$f_{18} = \frac{\sum_{(x,y)\in M_6\cup M_7\cup M_{10}\cup M_{11}} w_3(x,y)}{\sum_{i=1}^{16}\sum_{(x,y)\in M_i} w_3(x,y)}, \quad (3.14)$$

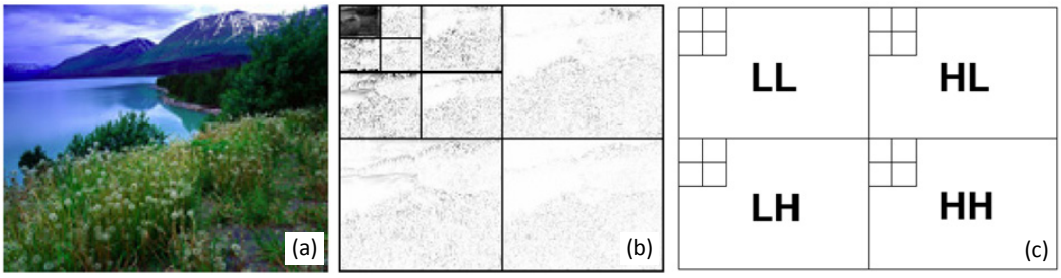


Figure 3.4. Daubechies wavelet transform (Datta et al., 2006). (a) Original image. (b) Three-level transform, levels separated by borders. (c) Arrangement of three bands LH, HL and HH of the coefficients.

f_{19} is being computed similarly for intensity I_V channel. We compute the rule of thirds feature for saturation as

$$f_{20} = \frac{9}{|I_S|} \sum_{x=X/3}^{2X/3} \sum_{y=Y/3}^{2Y/3} I_S(x,y), \tag{3.15}$$

where X and Y are the width and height of the saturation image I_S respectively. Rule of thirds for I_V, f_{21} is computed in a similar manner.

Camera settings information can be obtained from the EXIF data. However, EXIF data is not implicit to an image and not readily available for all images. Therefore, for practicality, we do not consider features related to camera settings.

3.1.2.2 Features of Salient Regions

We compute the measures of exposure, sharpness, and texture details for the salient regions utilizing the same respective techniques used to compute the global features. The exposure feature for the salient region is

$$f_{22} = \frac{1}{|S_V|} \sum_{m} S_V(m), \tag{3.16}$$

where m represent the pixels of the salient region. The saturation for the region, f_{23} is similarly computed. The sharpness features, f_{24} to f_{26}, are computed by applying our enhanced sharpness detection technique on the salient regions of the image. For texture details, we only compute the sum of the average wavelet coefficients over all levels to produce f_{27} to f_{29} for each HSV channel of the salient regions. Another photographic rule, *fill the frame*, suggests that the subject should occupy a large portion of the image. We represent the size of the salient regions by the dimension of the salient regions, and we have $f_{30} = |S_{ch}| = M$.

In addition to the features of the salient regions, we analyze the position, distribution, and the total number of salient locations. A professional photo has a strong focus and the subject can be easily identified. This feature is characterized by a small number and dense distribution of salient locations. We let $f_{31} = |L|$, where L is the number of unique salient locations of the image. To represent the distribution of the salient locations, we compute the standard deviation of all the salient locations, f_{32}.

A saliency map provides additional useful information about the salient regions. In addition to just the locations, the intensity and size of the salient regions represent the degrees of saliency of the corresponding salient regions. We compute the saliency map mean,

$$f_{33} = \frac{1}{|SM|}\sum_x\sum_y SM(x,y), \tag{3.17}$$

and the standard deviation,

$$f_{34} = \sqrt{\frac{1}{|SM|}\sum_x\sum_y\left(SM(x,y) - f_{33}\right)^2}, \tag{3.18}$$

to capture the saliency strength information. Comparing the two images in Figure 3.5, with aesthetics score of 6.56 and 3.83 respectively, image (a) has a total of 7 salient locations compared to 10 in image (b). In addition, image (a) yields smaller scores of 5.6 and 33 for f_{33} and f_{34} compared to scores of 7.8 and 38 obtained for image (b).

3.1.2.3 Features Depicting Subject-Background Relationship

From the survey conducted by Yan et al. (2006), they concluded that simplicity is the most distinguishing factor of professional photos. They used two *global* image

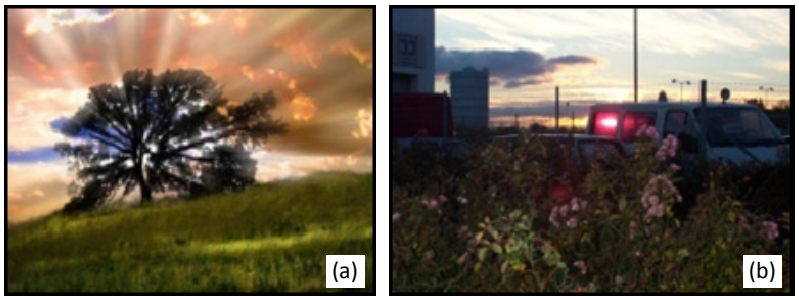


Figure 3.5. Effect of visual saliency of the photo subject on image aesthetics. (a) Professional photo with a prominent subject; aesthetics score = 6.46. (b) Snapshot without any prominent subject; aesthetics score = 3.83.

features—the edge spatial distribution and hue count—to measure the simplicity factor. As simplicity of a photo is mostly characterized by a simple background as well as clear contrast between the subject and the background, it would be more intuitive to compute simplicity measure based on the differences between the subject and its background in a number of aspects, such as the exposure, saturation, hue, blurriness, texture details, and edge spatial distribution. For example, difference in hue between the subject and its background represents the color contrast, which is an important photographic rule.

For these set of features, we apply the same methods used for the global image features to compute exposure, saturation, hue, blurriness, and texture details, for both the salient and background regions. We compute the differences of the respective features of the subject and its background using squared difference. For example, the subject-background difference for exposure is

$$f_{35} = \left(\frac{1}{|S_{\mathrm{V}}|} \sum_{m} S_{\mathrm{V}}(m) - \frac{1}{|B_{\mathrm{V}}|} \sum_{n} S_{\mathrm{V}}(n) \right)^2, \tag{3.19}$$

where m and n represent the pixels of the salient and background regions. Similarly, f_{36} and f_{37} are computed for subject-background differences in saturation and hue

respectively. The sharpness of the subject and background are computed by applying our enhanced sharpness detection technique to both the salient and background regions separately. The subject-background sharpness difference features, f_{38} to f_{40}, are then computed using the squared difference. The texture difference between the subject and background, f_{41} to f_{43}, are computed by taking the squared difference between the sum of the average wavelet coefficients over all three levels for each HSV channel of the salient regions and of the background region.

For edge simplicity, we first compute the edge spatial distribution (Yan et al. 2009) for the both the subject and background separately. We applied Canny edge detector to detect edges and compute the edge distribution by dividing the magnitude of the edge by the size of the salient regions and that of the background region. Then, the edge simplicity feature, f_{44}, is computed by calculating the squared difference between the edge spatial distributions of the subject and its background. Using images in Figure 3.6 for comparison, we obtained a set of consistently higher value for image (a) for all the subject-background features except the hue difference.

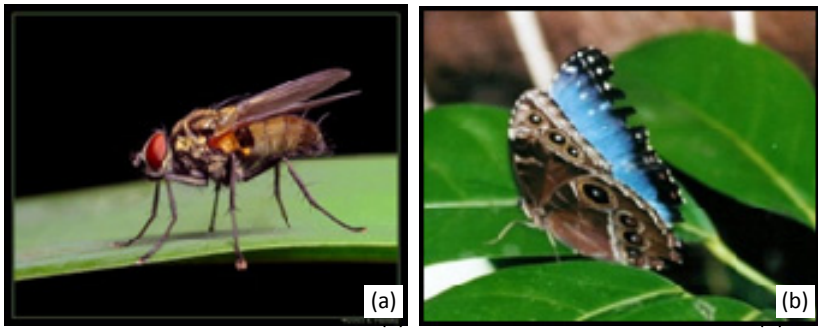


Figure 3.6. Effect of simplicity of the photo subject on image aesthetics. (a) Professional photo with a simple background, aesthetics score = 6.8. (b) Snapshot with a complex background, aesthetics score = 4.0.

This result is reasonable since the color contrast of image (a) is not very strong and the subject is made prominent by the other subject-background features.

3.1.3 Classification

To allow direct comparison with the results of Datta et al. (2006), we downloaded the same set of photos used by them from Photo.net. However, since some users have removed their photos from Photo.net, we managed to collect only a subset of 3161 photos out of the 3581 photos used by Datta et al. Each photo has an aesthetics score in the range of one to seven. The mean aesthetics score for this set of photos is 5.1. For training our classifier, we used only the top 10% and bottom 10% photos that have ratings above 6.2 and below 4.0 respectively.

Our feature set F consists of our global image features f_1 to f_{21}, and the salient features f_{22} to f_{44}. To illustrate the effectiveness of our salient region features, we have also created a feature set G by augmenting a selected set of Datta's most discriminative global image features with our set of salient features. Basically, G is a subset of F without our enhanced sharpness and contrast features where $G = F - \{f_1, f_2, f_3, f_4\}$. We perform attribute selection and classification on both feature sets using one-dimensional support vector machine (SVM) (Wang and Wiederhold, 2001) tool provided by Weka Explorer (Witten and Frank, 2005). We select SVM to build our model because it is a powerful binary classifier and is most appropriate for two-classification task. Instead of using SVM with RBF kernel (Datta el al. 2006), we chose to perform SVM classification without any kernel because we believe that professional and snapshots classes are linearly separable if our features are discriminative enough.

3.1.4 Experimental Results

Performing one-dimensional SVM on feature set G gives us the top 15 features with decreasing 5-CV accuracy, $G_S = \{f_{25}, f_{26}, f_8, f_{13}, f_7, f_{12}, f_{18}, f_7, f_5, f_{40}, f_{31}, f_{24}, f_{21}, f_{22}, f_{42}\}$ where $G_S \subset G$. Out of these 15 features, there are eight global image features and seven salient features. The top two features are our sharpness features for the salient region. This result shows that salient features do play an important role to differentiate professional photos from snapshots. We rerun attribute selection on F to obtain $F_S = \{f_2, f_3, f_8, f_{25}, f_4, f_{22}, f_{12}, f_7, f_{40}, f_{18}, f_{31}, f_{13}, f_7, f_5, f_{42}, f_{21}, f_{24}, f_{26}\}$ where $F_S \subset F$. Interestingly, three of our four enhanced global features representing image sharpness and contrast, emerged as top five features. Performing SVM classification on the top feature, f_2 produced an accuracy of 69.3%, demonstrating the high discriminative power of our sharpness features.

For classification, we obtained a 78.2% 5-CV accuracy for feature set G_S, with a high class precision of 82.9% and low class precision of 75.6%. For feature set F_S, the accuracy achieved is 78.8%, marginally higher than accuracy of feature set G_S. The precision for professional photos is increased to 83.7% but remains about the same at 75.2% for snapshots. This indicates that our enhanced sharpness and contrast global features are able to increase the discrimination of professional photos.

For comparison with Datta's approach, we run the SVM classifier using standard RBF kernel ($\gamma = 3.7$, cost = 1) on the top 15 features specified in their paper (Datta et al. 2006). Since our dataset is the subset of the original dataset used by Datta et al., for a fair comparison, we also perform SVM attribute selection and classification without RBF kernel on their full feature set. Figure 3.7 shows the comparison of our results with Datta et al.'s (2006) and Yan et al.'s (2006). Both of

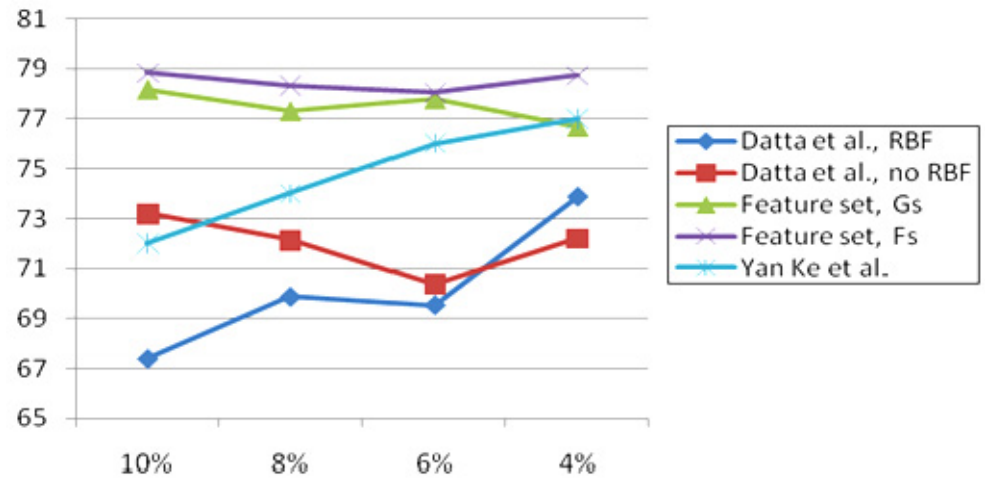


Figure 3.7. Comparison of classification accuracy with existing work on datasets consisting of the top and bottom n% of photos.

our feature sets, G_S and F_S outperformed existing work across the top and bottom n% datasets. Specifically, for the top and bottom 10% images, our 5-CV accuracy is about 6% to 8% higher than all existing works. Another interesting observation is that our results have much higher stability across the different datasets, maintaining an accuracy of 77% to 79%.

3.2 Aesthetics Score Prediction

Extending the aesthetics classification work above, we train a score prediction model to infer an aesthetics score for any given image. The inferred aesthetics score is in the range of 1 to 7, with a larger number being more aesthetically pleasing.

3.2.1 Salient Region, Visual Features Extraction and Regression

Figure 3.8 shows the three main steps in training an aesthetics score prediction model; subject extraction, feature extraction and regression. We first extract the salient region of an image using the same salient region extraction approach used in aesthetics class prediction, as described in Section 3.1.1. We then extract a set of 34 updated visual features, feature set *H* from each image. Feature set *H* consists of a subset of 29 features from feature set *F* (excluding the 15 wavelet-based texture features, f_6 to f_{19} and f_{27} to f_{29}) used in aesthetics class prediction and additional five new features—rule-of-thirds (Liu et al. 2010), visual balance (Liu et al. 2010), saliency of first subject, saliency of second subject, and saliency difference between first subject and background. This feature set can be

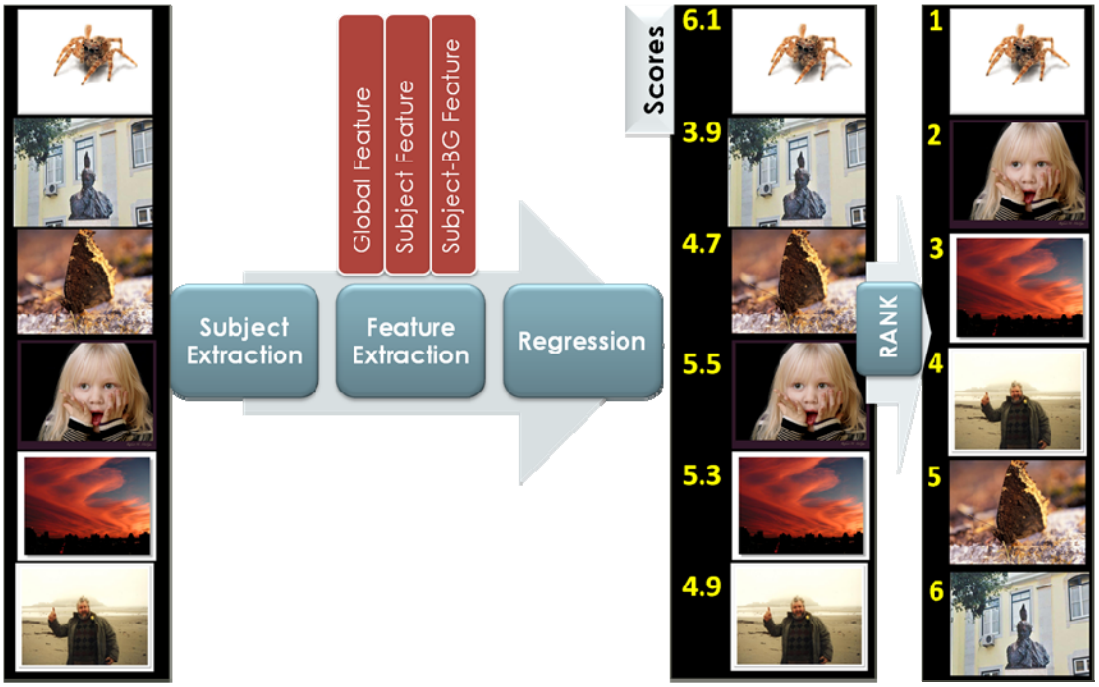


Figure 3.8. Overview of an aesthetics score prediction model.

categorized into two types – **low-level features** such as contrast, exposure, saturation, sharpness, and saliency and **photographic composition rules** such as the rule-of-thirds and visual balance.

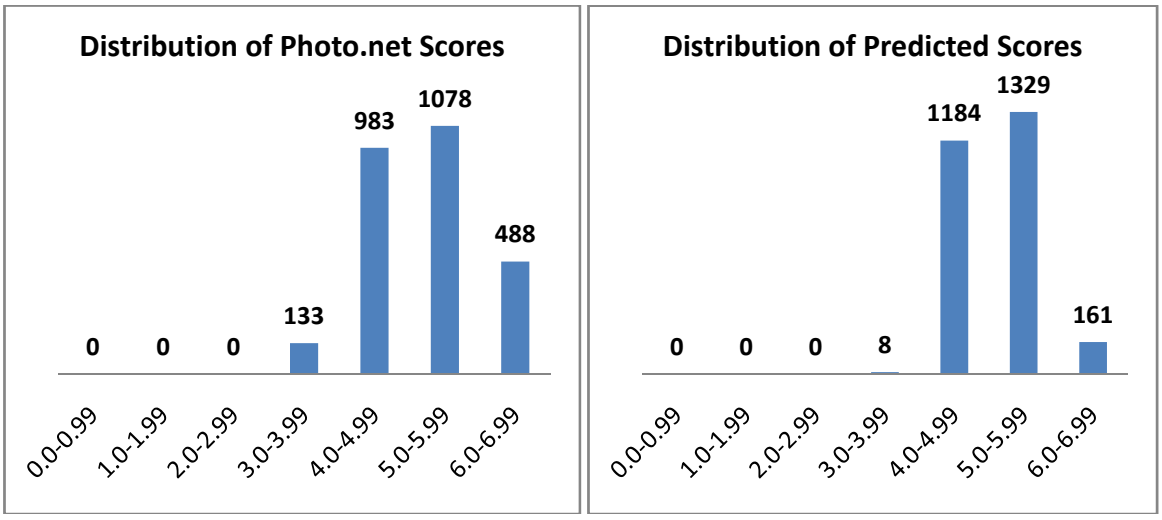
To train an aesthetics score prediction model, we performed 5th-degree polynomial regression on feature set $H = \{h_1, h_{2 \ldots} h_{34}\}$ to infer an aesthetics score in the range of 1 to 7. The aesthetics score model is formulated as

$$A = \sum_{i=1}^{34} (w_{i1} h_i + w_{i2} h_i^2 + w_{i3} h_i^2 + w_{i4} h_i^2 + w_{i5} h_i^5) \tag{3.20}$$

where $\{ w_{i1}, w_{i2}, w_{i3}, w_{i4}, w_{i5}\}$ represent the set of weights for the five polynomial terms of each feature h_i and $i = \{1,2 \ldots 34\}$. A set of 2682 images from Photo.net (Datta et al. 2006) is used in training the score prediction model. The distribution of the aesthetics scores of the Photo.net dataset is illustrated in Figure 3.9(a). We observe that the scores falls between the range of 3 to 7 and the average score is about 5.

3.2.2 Experimental Results

The regression produces a residual sum-of-squares $R_{res}^2 = 0.33$ and correlation coefficient value of 0.58. Figure 3.9(b) illustrates the distribution of the original scores from Photo.net and the scores predicted by our score prediction model. While the range of the predicted scores still falls within the range of 3 to 7, we observe that majority of the predicted scores are in the range of 4 to 6, with only a handful of the scores has a value below 4.0. Figure 3.10 shows some images with scores predicted by our model. Interestingly, results show that the photo subjects in images with higher aesthetics score appear significantly more outstanding compared to images with lower score.



(a) (b)

Figure 3.9. (a) Distribution of original scores used for training the score prediction model. (b) Distribution of the aesthetics scores predicted by our score prediction model.

The correlation coefficient value of 0.58 is considered moderately high and is found to be of practical use for ranking images. In our saliency retargeting algorithm presented in Chapter 4, we employ this score prediction model to maximize the aesthetics of the resulting image. User evaluation demonstrates that our score prediction model is effective in comparing aesthetics of images

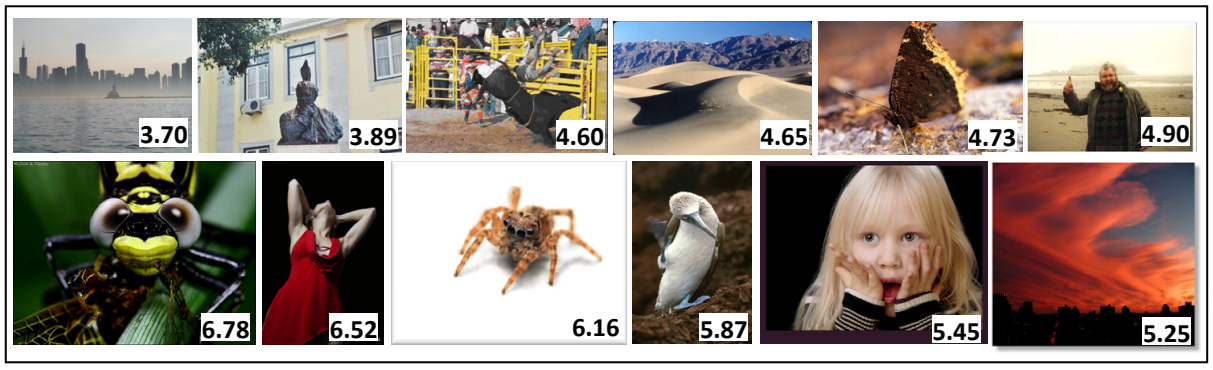
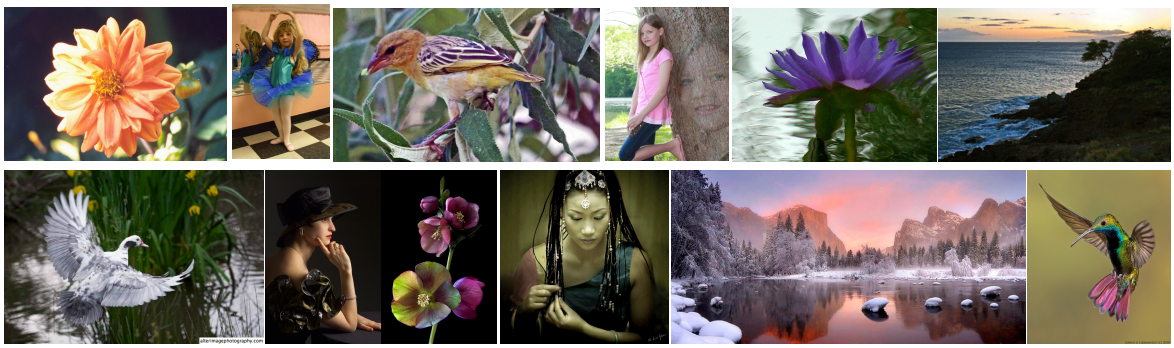
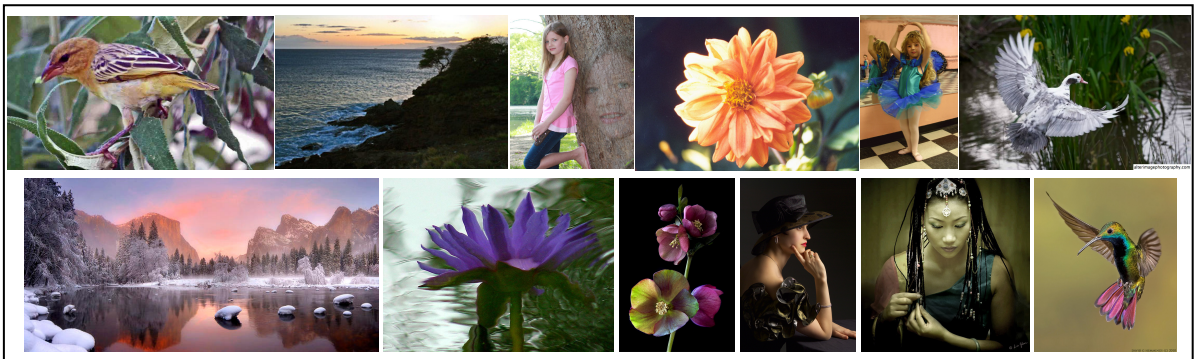


Figure 3.10. Images in ascending order of predicted scores (shown at the bottom right corner of each image).

before and after the image enhancement. Figure 3.11 compares the ranking of images generated by Photo.net and our score prediction model. Although our model does not produce absolute match with the Photo.net ranking in terms of ranking order, we find there is significant correlation between the two rankings. We can observe that majority of the images with lower scores that populate the first row of the Photo.net ranking remain in the first row of the ranking of our model. Vice versa, images with higher scores mostly retain their position in the second row. Further analysis reveals that only the two landscape images moves



Ranking of images based on **Photo.net scores** in ascending order.



Ranking of images based on **predicted scores** in ascending order.

Figure 3.11. Comparison of image ranking produced by photo.net and our score prediction model.

more than three positions in the ranking produced by our model. Most of the landscape images do not have an obvious photo subject and it is therefore not surprising that our saliency-based model does not produce accurate ranking for landscape images. This finding suggests that a category-based approach, particularly for landscape images can be adopted to further improves the performance of aesthetics score prediction.

3.3 Limitation and Future Work

Results show that our saliency-enhanced approach is indeed a promising direction for aesthetics class and score prediction. Our higher-level approach to extract salient region features proves to be more effective than the low-level region-based approach used by Datta et al. (2006). This result is not surprising since aesthetically-pleasing photos tend to direct the focus of attention to the intended subject that normally coincides with the salient locations. Furthermore, in professional photos, the subject normally stands out from the background. Representing the subject with the extracted salient regions and using it to identify the subject-background relationship enable the determination of the level of conspicuity of the subject. All this distinctive information of professional photos can be found in the saliency map as well as in the features extracted based on the salient regions, contributing to significantly higher classification accuracy. The consistency and stability of the performance of both our feature sets across the different image datasets illustrate that our saliency approach is less prone to misclassification due to noise inherent in individual image features.

Despite the promising result of our work, we believe that there are various limitations that can be addressed to further enhance the classification performance. Our salient region extraction relies on saliency model and image segmentation, both being open problems with results still not truly accord with human perception. There are possibilities in which the extracted salient regions may not represent, or only partially represent the photo subject, causing decreased classification and regression performance. Thus, future improvement in both saliency model and segmentation techniques would likely lead to better classification accuracy. Another limitation of our current work is that the number of salient locations used as seeds to extract regions from a segmented image is fixed to three. There are possibilities that these three seeds may not coincide with the subject, or in cases where the subject has high level of texture details as in low DOF macro images, three seeds are likely not sufficient to fully segment out the entire subject. Thus, finding the optimal number of seeds to be used for salient region segmentation may potentially lead to better performance. Additionally, an area worth investigating is the relationship among multiple salient regions for discovery of discriminative features.

In addition, we will also look into combining our saliency approach with a category-based approach. Different categories of photos have different set of features to determine whether they are good. For example, low DOF is one of the significant features that make good portrait photographs but it is not for landscape photographs. Vice versa, rule of thirds is more significant for landscape photographs than to portraits. Thus, combining both approaches may be a key for better classification and score prediction performance.

3.4 Chapter Summary

We have introduced a saliency-optimized approach for aesthetics class and score prediction. We identify the salient locations of an image using a visual attention model and use these salient locations to extract segments of image that coincides with these locations. Then based on the saliency map and the extracted salient region, we extract a set of salient features that reflects the characteristics of the salient region and depicts the subject-background relationship. Combining these set of features with a set of discriminative global image features, we achieved classification accuracy of 78.8%, which is considered to be significantly higher than existing work. For score prediction, despite achieving a moderate correlation coefficient value of 0.58, the model is found to of practical use. Employing this model to guide the aesthetics maximization of our saliency retargeting algorithm in Chapter 4, we find that this model is quite accurate in computing the relative aesthetics scores for the input and edited images. These results show that visual attention and aesthetics are indeed correlated, and our direction of employing saliency features for aesthetics prediction is promising.

Chapter 4

Saliency Retargeting: Aesthetics-driven Low Level Image Enhancement

In the photographs themselves there's a definite contrast between the figures and the location.

Helmut Newton

In this work, we develop the **saliency retargeting** algorithm, a novel low-level image enhancement method that alters image features of the objects in the photograph to match the order of visual importance intended by users. The goal is redirect the viewers' attention to the most important objects in the image, with the aim to produce more aesthetically pleasing resulting images. The primary application of this new approach is to enable casual photographers and novice photo-editing users to more easily improve their photographs. Casual

photographers often take photographs that do not have clear subjects or have the wrong objects being the subjects. They are generally aware that the photographs are not good, but they often do not know what is wrong and do not know how to enhance the photographs through digital photo editing. However, they know very well which objects are the intended subjects in the photographs, and this information is all that is required by saliency retargeting to enhance the photographs. For more advanced users, they can use our approach for advanced image editing that involves editing of sub-parts of objects in the image. This may result in many object segments being chosen and it is non-trivial for users to find the right combination of image modifications of the object segments to achieve the desired saliency. The saliency retargeting approach becomes even more useful for batch image enhancements

4.1 Approach

For an image to be enhanced using saliency retargeting approach, the user first has to create a segmentation map to partition the image into object segments. Fully automatic segmentation is not appropriate in this case because the input images may have non-salient main subjects, causing automatic segmentation to have high failure rates in identifying the object segments. Thus, we provide an interactive interface that incorporates GrabCut (Rother et al. 2004) for user to make trivial selection to partition the image into object segments. Next, the user specifies the *order of importance* for the object segments, where internally the object segments will be assigned unique *importance values* that accord with the order of importance.

With the above inputs, our approach then modifies the input image to produce

an enhanced image that satisfies the specified order of importance and has increased aesthetics. Our approach consists of two major components—a *saliency retargeting process* that modifies low-level image features of an image such that the computed saliency become consistent with the user-intended order of visual importance, and an *aesthetics maximization algorithm* that generates a set of images that satisfies the user-intended order of importance using the saliency retargeting process and returns the enhanced image with the highest score.

The core of the *saliency retargeting process* is an optimization algorithm that finds a set of image modifications to produce an image such that the saliency measurements of the object segments "match" their target importance values. In this work, only three low-level features in each object segment are modified,

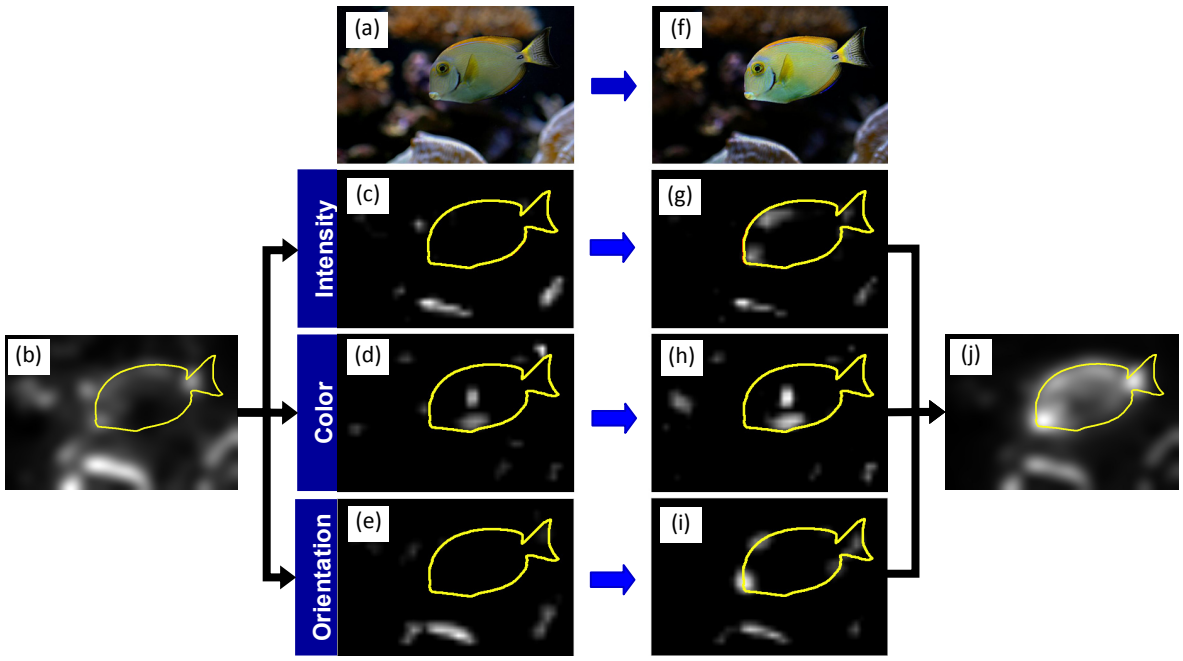


Figure 4.1. Effects of the image modifications on the conspicuity maps. (a) Input image, (b) its saliency map, and (c)-(e) the corresponding thresholded conspicuity maps. (f) Enhanced image produced by our saliency retargeting algorithm, (j) its saliency map, and (g)-(h) the corresponding thresholded conspicuity maps.

namely, luminance, color saturation, and sharpness. The reason for the choice of these features is that they correspond directly to the main features used in the widely-used Itti-Koch saliency model (1998), which we use to compute the image saliency. Figure 4.1 shows the effects of the image modifications on the conspicuity maps that collectively form the saliency map in Figure 1.3. This example shows that the saliency retargeting algorithm has successfully altered the intensity, color contrast and orientation conspicuity maps through local modifications to the luminance, color saturation and sharpness respectively for each object segment, leading to the increase in dominance of the fish.

This method also supports the change of visual importance of sub-parts of objects. For example, the features of the face in Figure 4.2(a) look flat due to unfavorable lighting. Based on the given order of importance of the regions marked by the mask in 4.2(c), the saliency retargeting algorithm retargets the saliency of image 4.2(a) to produce image 4.2(b), making the facial features more dominant.

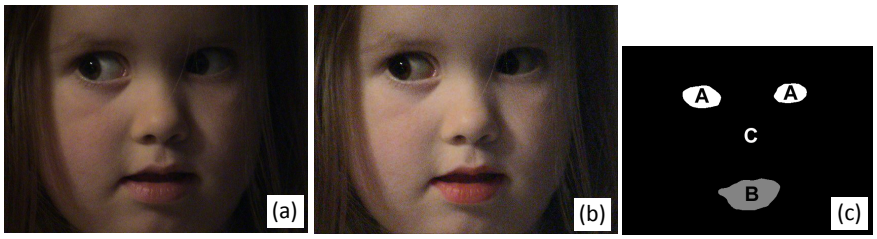


Figure 4.2. (a) Original image. (b) Enhanced image based on the mask in (c) with the order of importance given to the segments.

Due to there are uncountable number of images that can satisfy the user-intended order of importance, we perform aesthetics maximization to return a maximally-aesthetics version of a saliency-retargeted image. To enable aesthetics maximization, we trained an aesthetics score prediction model as described in

Section 3.2 to measure image aesthetics. The overview of the aesthetics maximization algorithm is illustrated in Figure 4.3. The algorithm first generates a set of vectors where each contains a different sequence of importance values that accords with the user-specified order of importance for the object segments. Each vector of importance values is then passed to the saliency retargeting algorithm to generate an enhanced image in which the average saliency values of the object segments are consistent with the respective importance values. Finally, the aesthetics score prediction model computes an aesthetics score for each enhanced image in this resulting image set and return the maximally-aesthetics version as the result. Figure 4.4(c) and 4.4(d) shows two of the enhanced images produced by the saliency retargeting algorithm for different vectors of important values generated

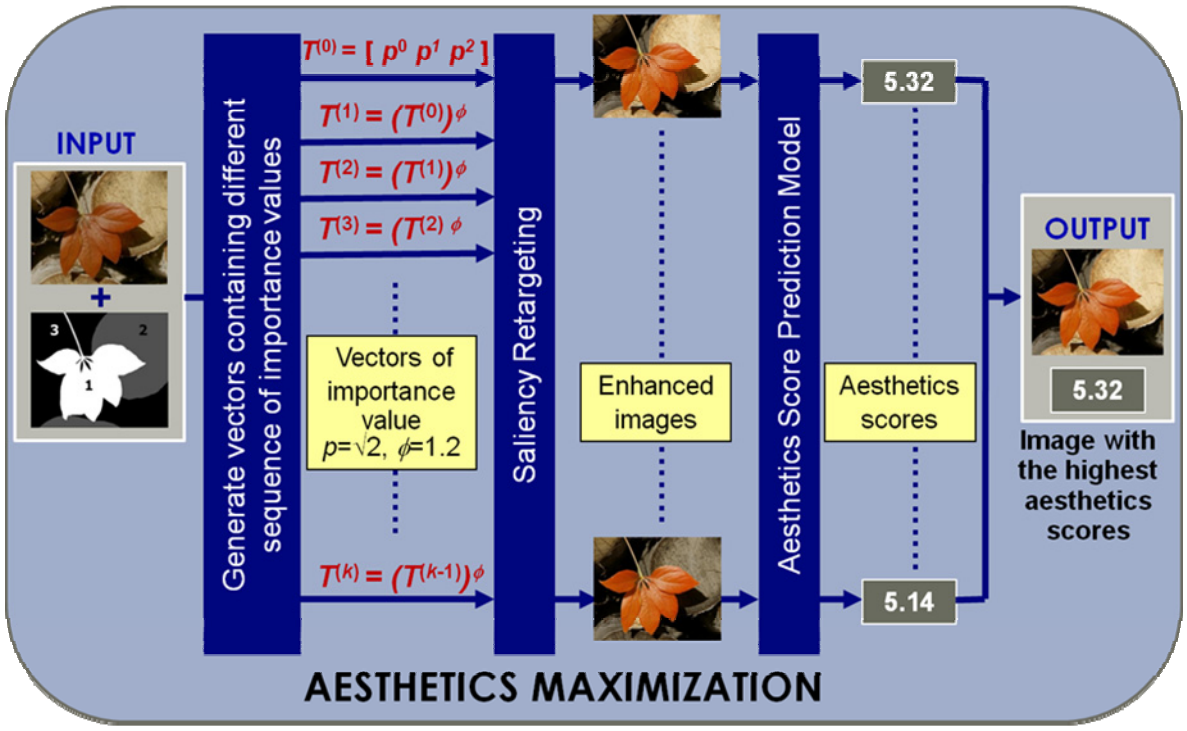


Figure 4.3. Overview of Aesthetics Maximization algorithm.

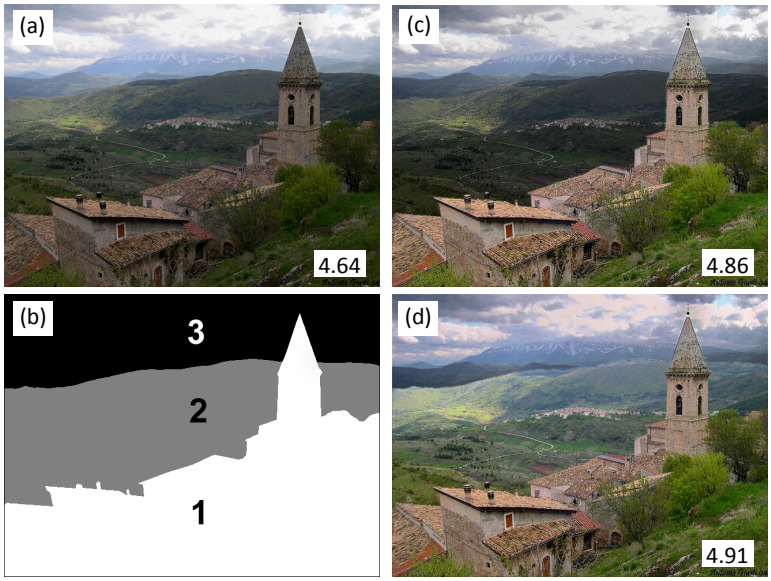


Figure 4.4. Example enhanced images (aesthetics score is at the bottom right of each image). (a) Original image. (b) Segmentation map with order of importance. (c)-(d) Two result images produced by our saliency retargeting algorithm for different sets of important values.

based on the user-specified importance map in Figure 4.4(b). Both enhanced images have higher aesthetics score than the original image in Figure 4.4(a). Image 4.4(d), the image with the highest aesthetics score is returned as the result enhanced image. Algorithmic details of our approach are described in the following sub-sections.

4.1.1 Saliency Retargeting

Given an image I and a set of N object segments S with target importance value T_i for each object segment i, we aim to enhance the aesthetics of the image by applying a set of low-level image modifications x to the input image I to produce an output image, O with saliency value M_i that "matches" the target importance value T_i for

each object segment i. We formulate the saliency retargeting process as a minimization problem as follows,

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{4.1}$$

where

$$f(\mathbf{x}) = \sum_{i=1}^{N} |q(T_i) - q(M_i)|, \tag{4.2}$$

$$M_i = S(P(I, \mathbf{x}), \tag{4.3}$$

object segment i=1 is the most important object segment while i=N is the least important object segment,

$\mathbf{x} = \{ v_i, s_i, \sigma_i \mid i = 1, 2, \ldots, N \}$ and $|\mathrm{x}| = 3\mathrm{N}$,

v_i is the increase of average luminance in object segment i,

s_i is the increase of average color saturation in object segment i,

σ_i is the increase of sharpness in object segment i (positive σ represents image sharpening, and negative σ image blurring),

P is a function that performs a set of image modifications $\mathbf{x}$ on input image I to produce an output image, O.

Q is a function that computes the average saliency for an object segment i in the output image and

$q(.)$ is a normalization function defined as

$$q(X_i) = \frac{X_i}{\sum_{j=1}^{N} X_j}. \tag{4.4}$$

Each region saliency M_i is the average saliency value within object segment i in the output image. A higher saliency value corresponds to higher saliency.

To avoid intensity and color saturation reversal, the image modifications $\{v_1, s_1\}$ for the most important object segment 1 are set to be positive. In addition, the constraint on the sharpness parameter σ_i is based on the assumption that the more

important object segment should be sharper than or as sharp as the less important object. Therefore, the optimization is subject to the set of inequality constraints

$$v_1 \geq 0,\ s_1 \geq 0, \text{and}$$
$$\sigma_i \geq \sigma_j \text{ iff } i \leq j. \tag{4.5}$$

In addition, a set of bound constraints is defined to ensure that the image is not over-modified resulting in image unnaturalness:

$$-v_B \leq v_i \leq v_B, -s_B \leq s_i \leq s_B, -\sigma_B \leq \sigma_i \leq \sigma_B. \tag{4.6}$$

4.1.1.1 Implementation

We employed Sequential Quadratic Programming (SQP) (Gill et al. 1981) to solve the saliency retargeting problem. Given the optimization problem defined in Eq. (4.1), SQP tries to solve the Lagrangian function

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j \cdot G_j(\mathbf{x}), \tag{4.7}$$

where G_j(x) is the combined set of inequality and bound constraints defined in Eq. (4.5) and (4.6).

The initial values of the optimization parameters $v_i^{(0)}$, $s_i^{(0)}$, and $\sigma_i^{(0)}$ is set to correlate with the difference between the saliency of the input image and its corresponding importance value:

$$v_i^{(0)}, s_i^{(0)}, \sigma_i^{(0)} = \begin{cases} 0.1 & q(T_i) \leq q(M_i) \\ -0.1 & q(T_i) > q(M_i) \end{cases}. \tag{4.8}$$

For bounds setting, the values; v_B = 0.15, s_B = 0.08 and σ_B = 0.2, produce reasonably good and natural results for most images.

4.1.1.2 Image modification

The set of optimized parameters $\mathbf{x} = \{v_i, s_i, \sigma_i\}$ is used to modify the input image. The modification is performed in the HSV color space. To obtain more natural images, it is important to consider the contrast sensitivity of human visual perception (Manos and Sakrison, 1974) and maintain the contrast within an image segment when adjusting its luminance. Therefore, we apply non-linear contrast-preserving changes to the V channel based on the perception theory that our eyes perceive higher contrast at the lower range of luminance compared to the higher range of luminance. For every pixel in object segment i, the new luminance value is computed as

$$V' = x + 6v_i(V^2 - V). \tag{4.9}$$

This formula ensures that the sum of luminance changes is the same as adding the value v_i directly to the V channel of each pixel in the object segment. Modification to the color saturation is performed by adding the parameter s_i directly to the S channel of the respective object segments. The sharpness effect is applied to the V channel after the luminance adjustment. The parameter σ_i is multiplied by a user-specified parameter γ such that $|\gamma\sigma_i|$ represents the sigma of the unsharp mask filter for sharpening or the standard deviation of the Gaussian filter for blurring. This γ parameter allows users to set the magnitude of the sharpness transformation to achieve their desired depth of field. In all experiments, $\gamma = 3$ is used for all images.

4.1.2 Aesthetics Maximization

The proposed aesthetics maximization algorithm first generate a set of vectors, each

containing a different sequence of importance values that satisfies the user-specified order of importance of the object segments. Each vector is then used by the saliency retargeting algorithm to produce an enhanced image in which the region saliency values of the object segments match the importance values in the vector. The aesthetics score prediction model described in Section 3.2 is then used to compute an aesthetics score for every enhanced image and the image with the highest score will be selected as the result image.

We set the initial vector of importance values as a geometric sequence where

$$T_i^{(0)} \leftarrow p^{N-i}, \tag{4.10}$$

for object segment i = 1, 2, ..., N. We then iteratively update this vector of importance values exponentially, i.e.

$$T_i^{(k+1)} \leftarrow (T_i^{(k)})^{\phi}, \tag{4.11}$$

to generate the set of vectors containing importance values. Each vector $(T_1^{(k)}, T_2^{(k)}, \dots, T_N^{(k)})$ is passed to the saliency retargeting algorithm to generate an enhanced image. Using the geometric sequence in Eq. (4.10) to generate the initial importance values signifies that the initial relative importance from one object segment to the next object segment is constant. The vector updating process in Eq. (4.11) gradually widens the relative difference in importance values between the object segments. Values of p and ϕ that are too small lead to generation of redundant enhanced images that are very similar, thus increasing computational time unnecessarily. On the other hand, values of p and ϕ that are too large may potentially skip important solutions. From the experiments, values $p = \sqrt{2}$ and $\phi = 1.2$ are found to be appropriate. To further reduce the search space, we limit the normalized importance value $q(T_i)$ to a maximum value of 0.85. Observation

shows that any object segment that carries an importance value more than 0.85 tends to produce over-dominance effect that often leads to reduced aesthetics. The reason for updating the importance value vector as described in Eq. (4.11) is to generate a set of enhanced images where the main subject becomes progressively more dominant.

Figure 4.5 depicts the aesthetics maximization algorithm where SALIENCY_RETARGET is a function that perform the optimization in Eq (4.1) to obtain a set of image modifications **x**, MODIFY_IMAGE is a function that apply the set of image modifications **x** on input image *I* to produce the output image *O* and COMPUTE_AESTHETICS_SCORE function compute the aesthetics score for output image *O* using the model described in Section 3.2.

```
AESTHETICS_MAXIMIZATION(I,S,N)
     p ← √2
     φ ← 1.1
     τ ← 0.85
     k ← 0
     for i=1,2,….N
          T_i^(0) ← p^(N−i)
     maxscore ← 0
     while (T_1^(k) < τ) do
          x = SALIENCY_RETARGET(I,S,N,T)
          O = MODIFY_IMAGE(I,S,N,x)
          score = COMPUTE_SCORE(O);
          if score > maxscore then
               score = maxscore
               maxO = O
          for i=1,2,….N
               T_i^(k+1) ← (T_i^(k))^φ
          k ← k + 1
     return maxO
```

Figure 4.5. Aesthetics Maximization algorithm

4.2 Experimental Results

4.2.1 Results

We tested the saliency retargeting algorithm on a set of test images selected from personal photo collections, from Photo.net, and from the BSDS segmentation benchmark image set (Wollen 1978). Each test image is segmented into 2 to 6 object segments, and ranked based on the importance of the object segments. It typically takes 30 to 90 seconds to segment an image. For 80% of the test images, the saliency retargeting algorithm is able to converge to solutions that produce retargeted images with improved aesthetics scores. Interestingly, for the 20% of the test images that result in retargeted images that do not have improved aesthetics scores, the human subject experiment shows that these retargeted images are not preferred over the originals by human viewers.

Figure 4.6 and 4.7 show some of the results produced by saliency retargeting. For the example in Figure 4.6, the average saliency values of the object segments in the input images do not correlate with the desired order of importance. In the

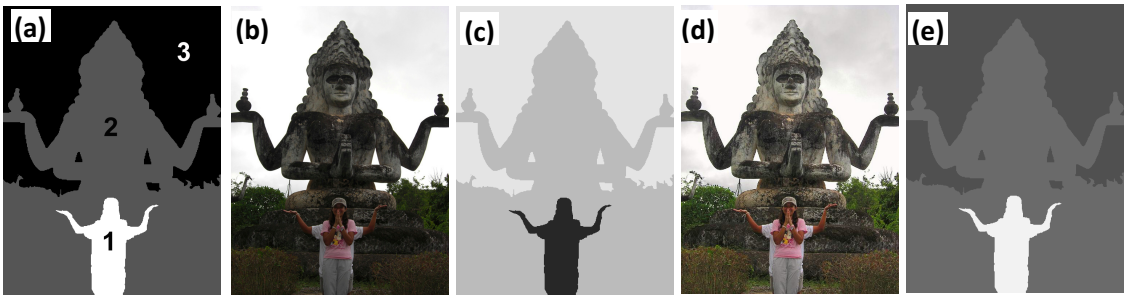


Figure 4.6. An example result. (a) Object segments and the desired importance order. (b) The input image. (c) The average saliency of each object segment in the input image. (d) The result image. (e) The average saliency of each object segment in the result image.

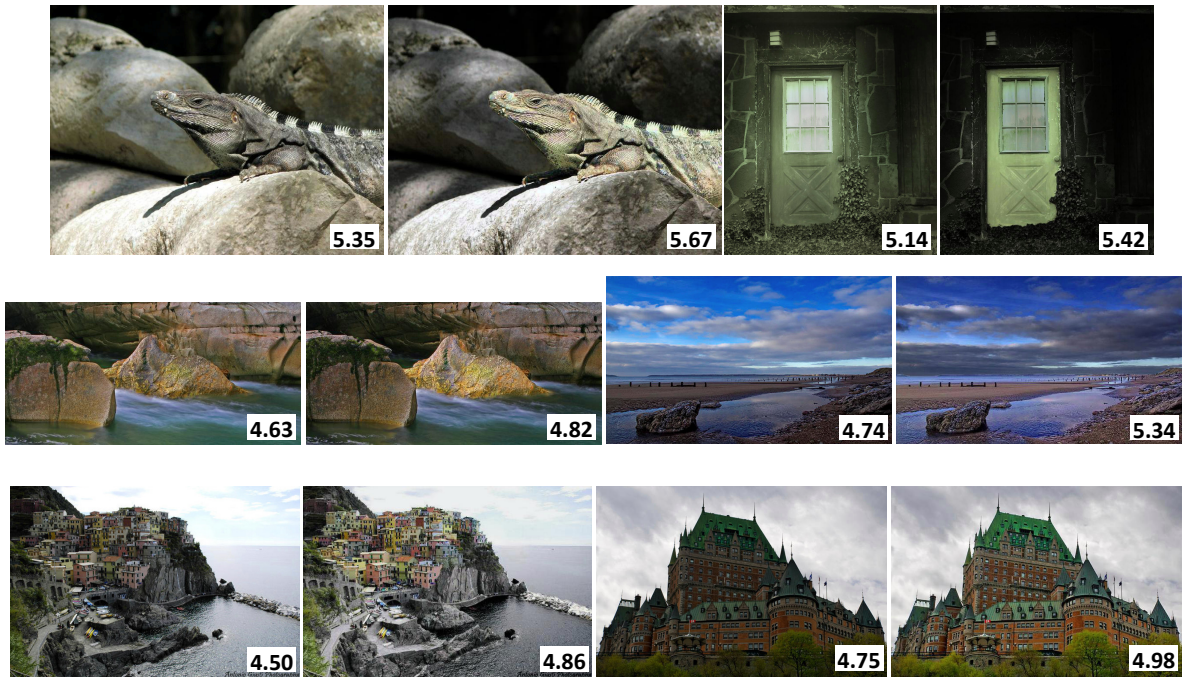


Figure 4.7. More results. For each pair of images, the one on the left is the original and the one on the right is the enhanced. At the bottom right is the computed aesthetics score for each image.

result, the algorithm has successfully altered the input image to achieve the targeted order of importance. It can be observed that the enhanced image is produced by local modification of the low-level features in each object segment. The computed aesthetics score of this test image increases from 4.59 to 4.94 after the enhancement.

4.2.2 User Evaluation

For an objective evaluation of the saliency retargeting approach, we conducted two human subject experiments to validate the effectiveness of our approach in

enhancing the subject dominance and image aesthetics respectively.

4.2.2.1 Validation of Subject Dominance Enhancement

In the first experiment, each test subject views either a set of 16 original images or a set of 16 images enhanced from the original images and his/her gaze is tracked while he/she is looking at each image. Each image is displayed for 3 seconds. The test subject does not know which set of images he/she is looking at. The goal of this experiment is to evaluate the effectiveness of the saliency retargeting method in modifying the actual visual saliency in the images. We recruited 28 subjects for this experiment.

Table 4.1 shows the correlation between the number of gaze fixations on each object segment and the desired importance value of the segment. One can see a significant increase in the correlation for the enhanced images. It shows that the saliency retargeting algorithm is quite effective in retargeting the actual saliency in the images. Figure 4.8 compares the scan paths on two original and their enhanced images. In both examples, the scan paths on the original images are diverse and less time is spent on the main subjects. On the other hand, on the enhanced images, the viewers' attention is concentrated around the main subjects, indicating that the viewers can more easily identify the main subjects in the image, which potentially leads to better aesthetics experience.

Table 4.1. Correlation between the number of gaze fixations and the desired importance value.

	Correlation coefficient	***p*-value**
Original images	0.284	0.0757
Enhanced images	0.5464	0.0003

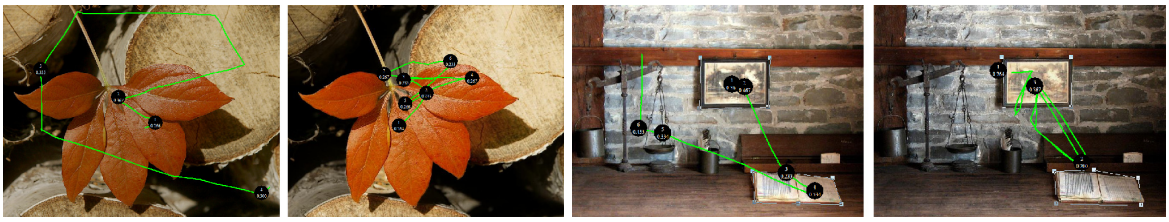


Figure 4.8. Scan paths on images. For each pair of images, the left one is the original and the right one is the result.

4.2.2.2 Validation of Aesthetics Enhancement

In the second experiment, we let each test subject compare 40 pairs of images. Each time, two images are shown side-by-side on the screen, where one is the original and the other has been enhanced by the saliency retargeting algorithm, and the test subject is required to choose one that he/she prefers. The screen positions (left or right) of the original image and its corresponding enhanced one are chosen at random and so the test subject does not know for sure which is the enhanced image. The purpose of this experiment is to study the effectiveness of saliency retargeting algorithm in enhancing the aesthetics of the images. The images we used in the user experiments are a subset of the test images mentioned in Section 4.2.1. We recruited 32 subjects for this experiment.

Results from this experiment clearly indicate that test subjects have a preference for the enhanced images. The chart in Figure 4.9(a) shows that for every image pair used in the experiment, at least 40% of the test subjects chose the enhanced image. For 87.5% of the image pairs, majority of the test subjects chose the saliency retargeted images. If we treat images with 40%–60% preference votes as noise, where there is no significant preference for the original or enhanced image, we can

conclude that the saliency retargeting algorithm is unlikely to produce images less aesthetically pleasing than the input images**.**

Next, we study the relationship between image aesthetics and the total saliency change. Aesthetics is represented by the ground truth of the number of preference votes an enhanced image gets. The total saliency change is computed by summing the saliency differences between the same object segments in the original image and in its corresponding enhanced image. The chart in Figure 4.9(b) shows two peaks, indicating there is high saliency change for two ranges of enhanced images. The total saliency change peaks at the enhanced images with the highest preference. The second peak falls within the range of 50%–59.9%, which can be considered noise. Studying the images within this range suggests that high saliency change in an image may potentially result in unnaturalness in the result, making it less likeable. In addition, computing the correlation between the number of preference votes an enhanced image gets and its total saliency change for images with more than 60% preference votes gives a coefficient of 0.42 (with *p*-value of 0.035). This suggests there is a significant

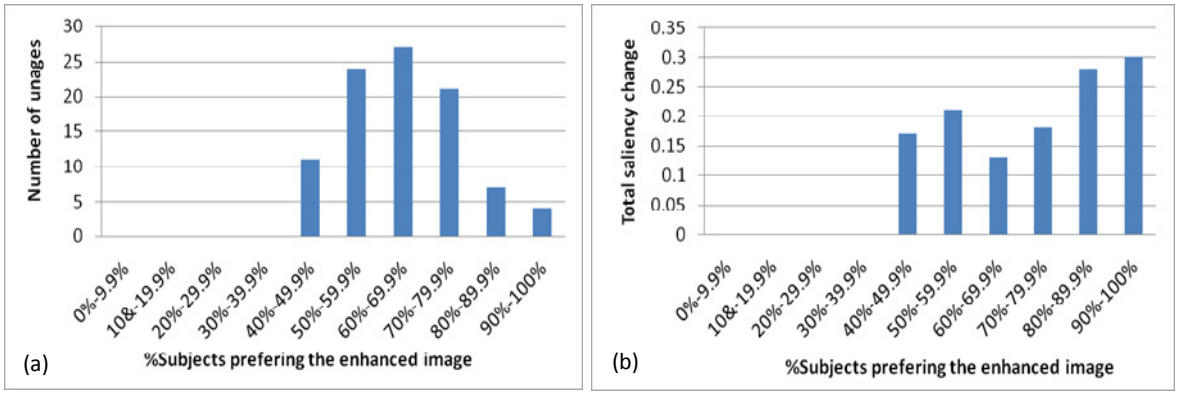


Figure 4.9. Results from experiment to validate aesthetics enhancement.

positive correlation between image aesthetics and saliency change.

To study the performance of the aesthetics score prediction model, we evaluate the accuracy of the aesthetics model in predicting the relative score between an original image and its corresponding enhanced images. For images with more than 60% preference votes, we find that the proposed model is able to predict an increase in aesthetics score with 88% accuracy. We notice that majority of the enhanced images that are given reduced scores by the aesthetics model fall under the category of images where consensus among human test subjects is low. This result is indeed encouraging, considering the subjectivity in evaluating aesthetics is unavoidable.

4.3 Limitation and Future Work

We are currently using fixed bounds for the amount of changes to the three low-level image features. We have observed that for some cases, these bounds are too conservative, and do not allow sufficient change of the image to satisfy the target saliency. On the other hand, the bounds appear to be too loose for some cases, and it results in overly-enhanced images that look unnatural as illustrated in the example results in Figure 4.10. Ideally, the image modifications should use image "naturalness" to limit the amount of changes; however, accurately evaluating the "naturalness" of images is still an open problem.

In general, the composition within a photograph is still one of the most important elements that determine the aesthetics of the image. Often, photographs are so poorly composed that no amount of saliency retargeting can make them more acceptable than before. Vice versa, in cases where main subjects are not salient,

Figure 4.10. (left) Input image. (right) Overly-enhanced image that appears unnatural.

re-composition alone cannot help to improve the image aesthetics much. Therefore, the aesthetics-driven re-composition method proposed in Chapter 4 can serve as good complement for saliency retargeting. These two image editing approaches could be integrated to produce a more complete content-aware image enhancement tool.

4.4 Chapter Summary

The main contribution of this work is the novel idea of using saliency retargeting as a means for image aesthetics enhancement, and a simple practical algorithm for the approach. The saliency retargeting is performed by modifying only the low-level image features that correspond directly to the features used in the saliency computation. Very importantly, the relationship between image aesthetics and image saliency, the goal of saliency retargeting, and the image features used for the saliency evaluation together provide a clear guidance to how the image can be modified to enhance its aesthetics. Results from user experiments have supported the effectiveness of saliency retargeting algorithm for enhancing image aesthetics.

Chapter 5

Saliency-based Image Recomposition and Image Retargeting

What's really important is to simplify. The work of most photographers would be improved immensely if they could do one thing: get rid of the extraneous. If you strive for simplicity, you are more likely to reach the viewer.

William Albert Allard

As explored in Section 2.1.1.1, many photographic rules are intended to enhance subject dominance, a highly desired aesthetics element. Among these rules, simplicity is the most important rule that many professional photographers are faithful to. In capturing stunning shots, professional photographers explore different camera position and view angle to find the choice of a simple background

that would make the subject more dominant. Despite the importance of the simplicity rule, none of the state-of-art recomposition approaches attempted to geometrically transform an image to enhance the visual dominance of a photo subject, largely due to the inability of all existing image geometric transformation operators to support the change in subject-background relationship without violating spatial semantics.

Motivated by the lack of such an image operator, we present tearable image warping, a semantics-preserving warping method that capacitate the change of the immediate background surrounding a photo subject. Tearable image warping has the ability to enhance visual dominance of the photo subject, by producing a natural change in the subject-background relationship, analogous to the change of viewpoint. In addition to enhancing subject dominance, this novel warping method can also be extended to support many other photographic rules such as rule-of-thirds, and visual balance. To our pleasant discovery, tearable image warping also works remarkably well for image retargeting, particularly in extreme resizing cases. In this section, we first present the conceptual overview and algorithmic details of tearable image warping. Following that, we provide the specific algorithmic details, implementation and results of employing tearable image warping in three applications; automatic image retargeting, semi-automatic image recomposition and interactive recomposition. Relevant empirical user studies have also been performed to validate the results. Comparison of our results with state-of-art approaches and results from the empirical evaluation demonstrate the effectiveness of tearable image warping in producing desirable results for both image retargeting and image recomposition.

5.1 Image Operator: Tearable Image Warping

We observed that in general, image topology does not need to be preserved everywhere to maintain semantics correctness. For example, the connectedness between an object's boundary and its adjacent background can be disregarded if it does not correspond to actual physical contact in the 3D world. On the other hand, the connectedness with the part of the environment where it comes into actual physical contact (e.g. the ground) should be preserved. In this dissertation, we introduce a new image warping method, named **tearable image warping**, that capitalizes on this idea for scene-consistent image editing.

5.1.1 Conceptual Overview

In addition to the two scene consistency properties defined by Mansfield et al. (2010) for image retargeting— zero object distortion and correct scene occlusion — our approach is able to further enhance scene consistency by preserving semantics connectness of objects with their environment. We thus redefine the scene consistency properties in the context of image editing as:

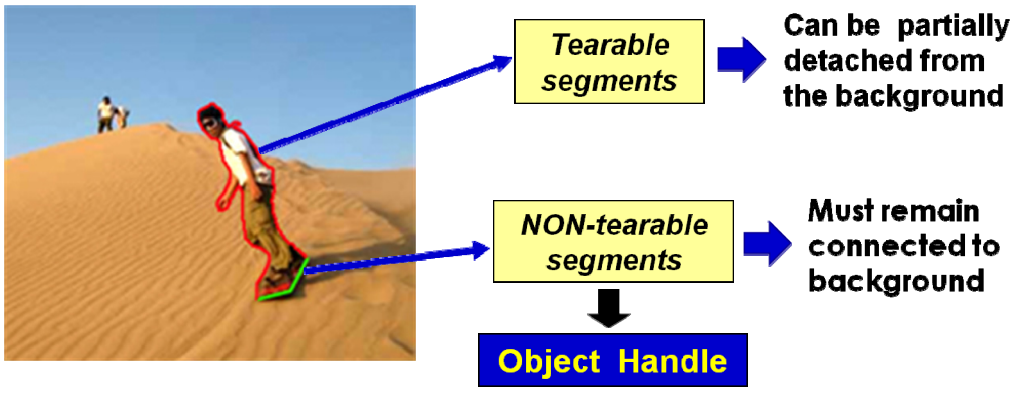
1) Objects are not distorted but kept as in the original image.
2) Objects are placed in their correct depth ordering.
3) Objects maintain consistent physical contacts with their environment.

To achieve zero object distortion and scene consistent occlusion, we adopt Mansfield's model decomposition approach, whereby an image can be described by a relative depth map comprising of object segments and their relative depth order.

To support the third property of scene consistency, we divide the boundary of each object into **tearable** and **non-tearable segments** as illustrated in Figure 5.1. Tearable segments represent boundary sections where depth discontinuity occurs and non-tearable segments correspond to boundary sections that have physical contact with environment and must therefore be preserved.

We use **object handles** to represent the non-tearable segments of an object. An object handle is a polyline drawn by the user to specify a part of an object's boundary that is non-tearable. In our implementation, an object handle marks a local area that has to be kept rigid, therefore the polyline does not need to be as precise as the object boundary. In general, an object handle can be anywhere in the object segment, and need not even be near the object boundary. For good image editing results, an object handle must satisfy two criteria:

- an object handle must be preserved as rigid as possible, ideally without rotation; and
- the object must be re-inserted to coincide precisely with its handle at a new location.



Fiigure 5.1. Conceptual overview of tearable image warping – tearable segments VS non-tearable segments.

Each object can have multiple object handles. Multiple handles of an object are combined as one object handle in the warping process. More examples of photo subjects with single and multiple handles are shown in Figure 5.5, Figure 5.7 and Figure 5.8.

To allow more flexibility, tearable image warping can be used for completely tearable or completely non-tearable object boundary. Completely tearable objects, such as birds flying in the sky, can have *no object handle* and its resulting tearable warping is equivalent to pure cut-and-paste. Vice versa, completely non-tearable objects, such as windows, can be defined with *full object handle* and its resulting tearable warping is equivalent to pure image warping. In short, tearable image warping is a unified approach that can smoothly transition from pure cut-and-paste to pure image warping.

5.1.2 Algorithm

Given an image with a set of object segments and their respective object handles, the tearable image warping algorithm performs the following three main steps: (1) decomposition, (2) warping and (3) image compositing. Figure 5.2 shows an overview of the algorithm steps and the intermediate inputs/outputs.

In the decomposition step, the image is first decomposed into a background layer and potentially multiple object layers. Holes left by the objects in the background layer are automatically inpainted. In the warping step, traditional non-homogeneous warping is applied only to the background layer, always keeping the object handles as rigid as possible. In addition to the core warping energy, application-specific energy and/or constraints are applied to drive different applications of image editing. In the image compositing step, the objects are pasted

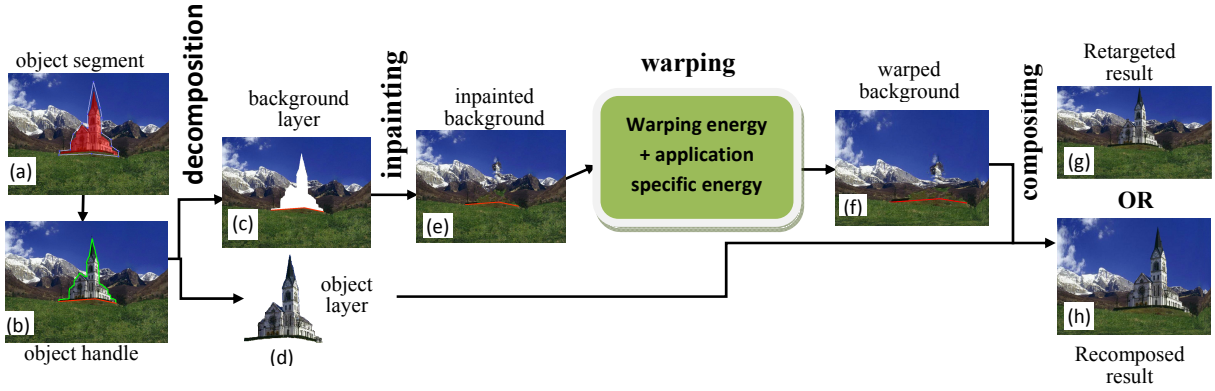


Figure 5.2. Steps in tearable image warping. In (b), the object boundary is shown in green and the object handle in red.

back to the warped background based on the new positions of their respective handles. Warping the background without the need to protect the entire objects gives more room to distribute the distortion more uniformly. The rigidity of the handles ensures that the objects can be seamlessly restored back to their original contact locations with the background.

5.1.2.1 Image Decomposition

The decomposition step uses the object segmentations supplied by the user to decompose the input image into a background layer and one or more object layers. A feathered mask is created for each object layer so that the object's boundary can be blended smoothly with the background when the object is re-inserted back to form the final image in the image compositing stage. Holes left by cut-out objects in the background layer can be automatically inpainted using any of the exemplar-based inpainting methods (Criminisi et al. 2004, Yousef and Hussien 2011).

5.1.2.2 Warping

The choice of a suitable base image warping method is important. Theoretically, any warping methods such as triangle-based, quad-based and pixel-based image warping methods can be adopted as long as it can support the preservation of object handles. We chose a triangle-mesh-based warping method for its efficiency and ease of representation. In a triangle mesh, an object handle can be easily made into edges of the triangle mesh using constrained Delaunay triangulation. Figure 5.3 shows an example initial triangle mesh for an input image, in which the green edges represent an object handle. In principle, any befitting mesh parameterization (Guo et al. 2009, Jin et al. 2010) can be used to find the destination locations for the triangle nodes. We adapted the non-homogenous scaling optimization method by Jin et al. (2010) to warp the background image, but without the saliency constraints and the weights to preserve salient content. Instead, we apply our *handle shape constraint* as a hard constraint to preserve the shapes and orientations of object handles. This method constrains the transformation of triangles to scaling and translation, without rotation, and is thus an ideal method to preserve object handles.

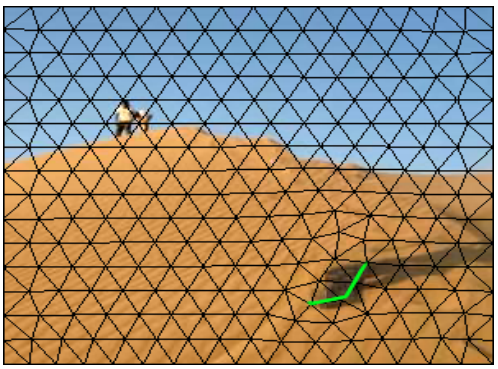


Figure 5.3. A triangle mesh used for warping. The object handle is highlighted in green.

Given a source mesh M for the input background image and the object handles, the warping process is the problem of mapping M to a target mesh M' that still keeps all the objects handles rigid. In addition to the core warping energies, there is a set of application-specific energies and/or constraints for each type of application. For image retargeting, M' must fit the target image aspect ratio. For automatic image recomposition, there is a set of aesthetics energies that drive the warping process to modify the spatial composition of an image to adhere to a set of photographic rules. After we have computed M', the new warped background image is obtained by an inverse piecewise affine mapping for each triangle in the mesh. The following subsections describe the core warping energy and constraints that are used in the computation of the target mesh M'. The application-specific energies and constraints for automatic retargeting, semi-automatic recomposition and interactive recomposition will be described in Section 5.2.1, 5.3.1.1 and 5.3.2.1 respectively.

5.1.2.2.1 Warping Energy

Scale transformation error. For each triangle $t \in M$, we constrain the transformation to non-uniform scaling (Jin el al. 2010), denoted by $G_t = \begin{pmatrix} s_t^x & 0 \\ 0 & s_t^y \end{pmatrix}$. However, in general, for each triangle in M, there is an affine mapping that maps it to its corresponding triangle in M' and the linear portion of the affine mapping can be represented by a 2×2 Jacobian matrix J_t. The scale transformation error is defined as

$$E_w = \sum_{t \in T} A_t \|J_t - G_t\|_F^2 , \tag{5.1}$$

where A_t is the area of triangle t in M' and $\|\cdot\|_F^2$ is the Frobenius norm.

Smoothness error. To avoid discontinuity in the resulting image, we enforce a smoothness term that tries to minimize the scale difference between neighboring triangles (Jin el al. 2010):

$$E_s = \sum_{\substack{s,t \in M \\ s \text{ and } t \text{ are adjacent}}} A_{st} \| G_t - G_s \|_F^2 , \tag{5.2}$$

where $A_{st} = (A_s + A_t)/2$.

Total error. The total warping energy is defined as the weighted sum of the scale transformation and smoothness errors:

$$E = \alpha E_w + \beta E_s , \tag{5.3}$$

where α and β are the weights. Minimizing the total error function will try to constrain the warping of all triangles to non-uniform scaling, without rotation. This total error, representing the core warping energy will be combined with the application-specific energy and/or constraints in order drive tearable image warping to perform the specific task such as image recomposition or image retargeting.

5.1.2.2.2 Handle Shape Constraint

To ensure that an object can be re-inserted seamlessly to its object handle in the warped background, we must preserve the shape and orientation of the handle during the warping process. Here, we assume that all object handles of an object has been combined into one. Suppose the object handle consists of n vertices, $v_1, v_2, \cdots, v_n$, in M, and they are being mapped to vertices $u_1, u_2, \cdots, u_n$ in M'.

To preserve the shape of an object handle, for each vertex v_i, we preserve two distance measures; (1) distance between v_i and v_1, and (2) distance between v_i and

v_{i+1}. To preserve the orientation of the object handle, each of the above distances is computed as a signed x-distance and a signed y-distance separately. Consequently, for each object k, the handle shape constraints are

$$\begin{aligned} u_{i,x} - u_{1,x} &= s_k(v_{i,x} - v_{1,x}), \\ u_{i,y} - u_{1,y} &= s_k(v_{i,y} - v_{1,y}), \end{aligned} \tag{5.4}$$

and

$$\begin{aligned} u_{j+1,x} - u_{j,x} &= s_k(v_{j+1,x} - v_{j,x}), \\ u_{j+1,y} - u_{j,y} &= s_k(v_{j+1,y} - v_{j,y}), \end{aligned} \tag{5.5}$$

where $i = 2, \cdots, n,\ \ j = 2, \cdots, n-1$, and s_k is a scale factor that allows the object handle (and the object) to undergo uniform scaling. For automatic image retargeting and interactive image recomposition, s_k is specified by the user or set to 1 by default. Each object can have a different scale factor. For automatic image recomposition, s_k is posed as an optimization parameter in order to achieve aesthetically pleasing size.

5.1.2.2.3 Boundary Positional Constraint

To keep the problem well-posed, we constrain the boundary vertices in the input mesh M to the boundary of the output mesh M'. For each vertex v on the left, right, top or bottom border of M', we apply the positional constraints $v_x = 0$, $v_x = W$, $v_y = 0$, and $v_y = H$, respectively, where W and H are the width and height of the output image, respectively.

5.1.2.3 Image Compositing

The warped background is combined with all the object layers to form the final output image in the image compositing step. Each cut-out object image is first

scaled by its respective scale factor s_k before it is re-inserted onto the warped background at its new object handle location. If object overlapping is allowed, we re-insert the object according to the given depth order.

5.2 Image Retargeting

Since image recomposition can also be applied on retargeted images, we first provide the algorithmic details, implementation and results of image retargeting in this section. We then provide the details for image recomposition in the Section 5.3, demonstrating some results of recomposition on retargeted images.

5.2.1 Retargeting-specific Constraints

Considering that image retargeting is an automatic process and objects are not warped together with the background, we chose to be more conservative in modifying the image. Therefore, we apply two additional hard constraints: (1) *object boundary constraint* to ensure that objects are not cropped off, and (2) *non-overlap constraint* to ensure that objects that do not overlap in the input image will not overlap in the output image. However, if a depth order is given, our algorithm can relax the non-overlap constraint to allow occlusion of objects by re-inserting the objects in the final image compositing step based on the depth order.

Object boundary constraint. We first compute an axis-aligned bounding box around the object. During optimization, we compute the new position of the bounding box based on the new position of the object handle. To enforce this

constraint, we disallow any part of the object's bounding box to move outside the target image region.

Non-overlap constraint. This is similar to the object boundary constraint, but we keep track of the bounding boxes of all objects and enforce that they do not overlap. Using axis-aligned bounding boxes for overlap testing is efficient but not accurate, since it may give false hits when the objects are actually not overlapping. For higher accuracy, complex polygons or hierarchical bounding boxes can be used. Figure 5.4 compares some results of tearable image warping with and without the non-overlap constraint.

5.2.2 Implementation

User Input and Interaction. The primary user input includes object segments and their respective object handles. We use GrabCut (Rother et al. 2004) to allow the user to segment the objects easily. Very often, the user only needs to draw a polygon around the object with a few clicks. Similarly, the users can specify a

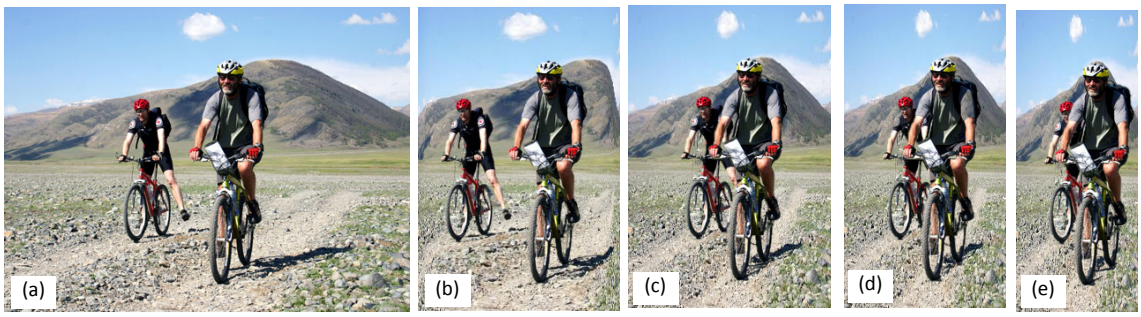


Figure 5.4. Image retargeting with and without the non-overlap constraint. (a) Input image, (b) retargeted with non-overlap constraint, (c)–(e) retargeted without the non-overlap constraint.

polyline to define an object handle. In addition, for retargeting-specific input, the user is prompted to provide the target scale factors s_x and s_y for the x and y directions respectively. If occlusion is toggled on, the user can choose to use the default depth order derived from the order of object segmentation or to provide the depth order for the objects.

Optimization details. To retarget an image, we utilize the CVX Matlab toolbox (Rother et al.) to find the solution to the convex quadratic function defined by the energy function in Eq. (5.3). The energy weights α and β are set to 1 and 0.5 respectively. The handle shape and image boundary constraints are set as hard constraints. The object boundary and optional non-overlap constraints are set as inequality constraints.

5.2.3 Results and Discussion

Results in this dissertation were generated on a laptop with Intel Core2 Duo CPU 2.53GHz and 4GB memory. Inpainting of the background image is performed in the pre-processing stage. It takes less than 10s to inpaint the area covered by the person in the leftmost image of size 664x1024 in Figure 5.5 using both the CPU-based (Teorex) and GPU accelerated (Yousef and Hussien 2011) inpainting methods. Excluding the time taken for inpainting, our tearable image warping algorithm produces a retargeted result in about 2s to 4s for an image of resolution 1024 x 768. This is a significant speed up compared to scene carving (Mansfield et al. 2010), which takes almost 27 mins on the same computer to retarget the same image to half of its original height. The speed of our algorithm depends mainly on the number of the triangle meshes per image, which we have kept almost constant even

for different image sizes. In contrast, the speed of scene carving decreases with increasing image size and increasing difference between input and output image sizes.

We compare our tearable warping method with a traditional warping method based on (Liu et al. 2010) and with scene carving (Mansfield et al., 2010). For fair comparison, we use the same manually segmented objects as ROIs for all methods. For the comparisons with traditional warping, we set the scale factor of all objects to max(sx, sy). To minimize object distortion for traditional warping, we apply hard constraint to preserve the salient triangles representing the objects. However, in extreme retargeting cases where no solution can be obtained for traditional warping, we relax the salient triangles preservation as soft constraint. For comparison with scene carving that allows objects to be cropped, we relax the object boundary constraint of tearable warping to allow objects to be cropped.

Figure 5.5 compares the retargeted results of tearable image warping with two state-of-art retargeting methods, traditional image warping (Liu et al. 2010) and scene carving (Mansfield et al., 2010). These examples are considered somewhat extreme retargeting because of their image contents and the target aspect ratios. To achieve the target aspect ratios, severe distortions to background, caused by over-compression often occur in the traditional warping method, as illustrated in its results in Figure 5.5 (third row). Results of traditional warping in column two and three of Figure 5.5 also exhibit obvious distortion to the main objects, as highlighted by the blue rectangles. In comparison, tearable image warping consistently distorts the background less in cases of extreme retargeting, as shown in its results in Figure 5.5 (second row). The retargeted triangle meshes shown in Figure 5.6 give some insights to how tearable warping can reduce the over-compression problem. In

Figure 5.5. (Top row) Input images with object handles shown in red, (second row) results of tearable image warping, (third row) results of traditional warping (Liu et al. 2010) and (bottom row) results of scene carving (Mansfield et al. 2010). Red rectangles highlight background distortion and blue rectangles highlight object distortion.

traditional warping, all triangles representing the object need to be preserved uniformly, leaving little room for distributing the compression. In tearable image warping, only edges representing the object handles need to be preserved, therefore compression can be distributed more evenly throughout the image, including areas "behind" the objects.

On the other hand, scene carving has little distortion in images with large homogenous regions, but for structurally complex images, severe distortions may

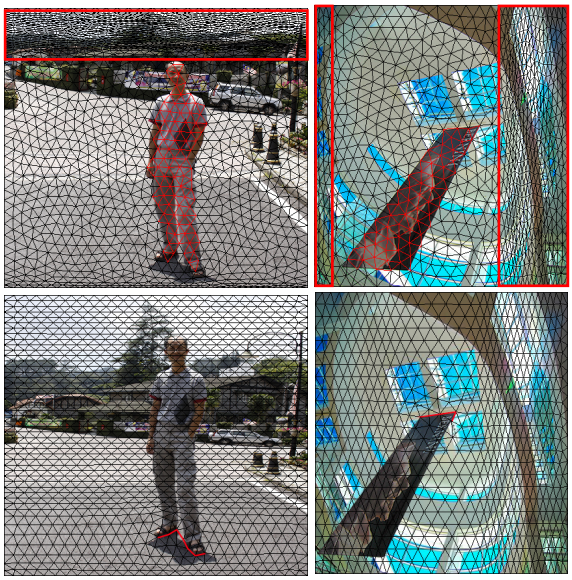


Figure 5.6. (Top row) Resulting triangle meshes of traditional image warping; (bottom row) of tearable image warping.

occur, as shown in its results in the last column of Figure 5.7. Noticeably, scene carving may result in cropping, thus potentially destroying the global context. In contrast, tearable warping can preserve the global context much better. A unique feature of tearable warping is the effect of objects shifting with respect to their immediate background. This effect can be observed in many of our results in Figure 5.5. Interestingly, this change of object-background relationship often produces a natural and semantics-preserving effect analogous to a shift of the viewpoint. This feature potentially is a powerful technique for aesthetics enhancement of images.

In terms of scene consistency, object distortion can never occur in tearable warping and scene carving because objects are not involved in the warping and seam carving process respectively. Furthermore, tearable warping and scene carving allow objects to overlap while maintaining the correct depth order, which gives more flexibility to perform extreme image retargeting without object

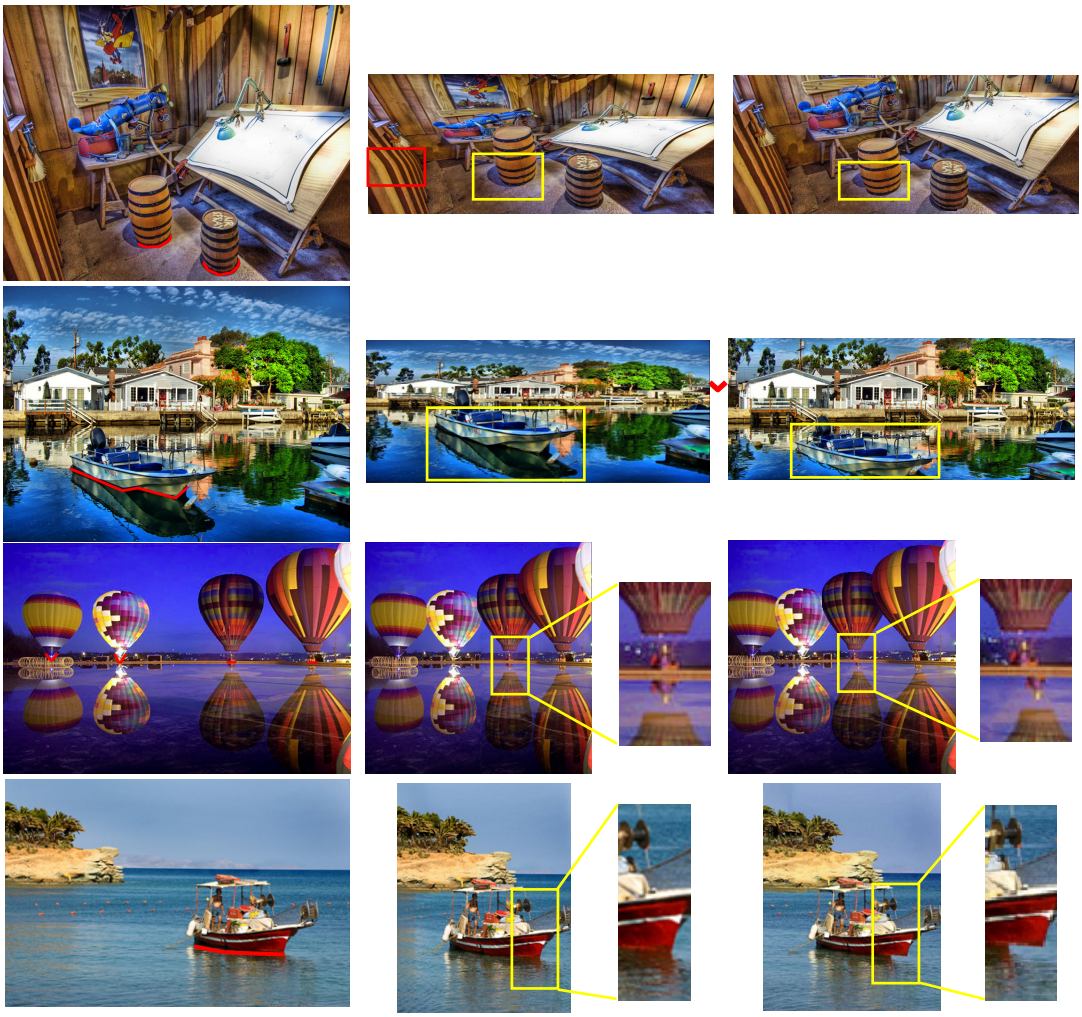


Figure 5.7. (Top row) Input images, with object handles shown in red, (middle row) results of tearable warping, and (bottom row) results of scene carving. Yellow rectangles highlight and compare the results of both methods in maintaining consistent physical contact between objects and their environments. The red rectangle highlights feature distortion.

distortion, as illustrated in the examples in Figures 5.8 and 5.9. In addition to its ability to better preserve global context, another key advantage of tearable warping over scene carving is its capability to maintain consistent connectedness of objects with their environments. Figure 5.7 shows examples in which scene carving fails in

this aspect, particularly when shadows, reflections or ripples exist, leading to unpleasant image artifacts. In contrast, tearable warping is able to maintain consistent physical contacts.

Tearable image warping maintains consistent semantic connectedness of an object by keeping the object handle rigid. To keep an object handle rigid, the handle must be defined by at least three non-collinear points. However, in cases where there are many cluttered objects, keeping all the object handles rigid will restrict overlapping of objects and thus forbid extreme retargeting. In such case, the handle can be relaxed by specifying only two points. With only two points, the handle will not be preserved rigidly but the relative positions between objects will still be maintained and more overlapping is thus allowed in extreme retargeting. This is demonstrated in the middle example in Figure 5.8 and the penguin example in Figure 5.9. Relaxed handles can also be used for cases where objects are not in physical contact with their environment, so that their relative positions could still be

Figure 5.8. Results with object occlusion. (Top) Input images with object handles shown in red, (bottom left) results of tearable warping, and (bottom right) results of scene carving.

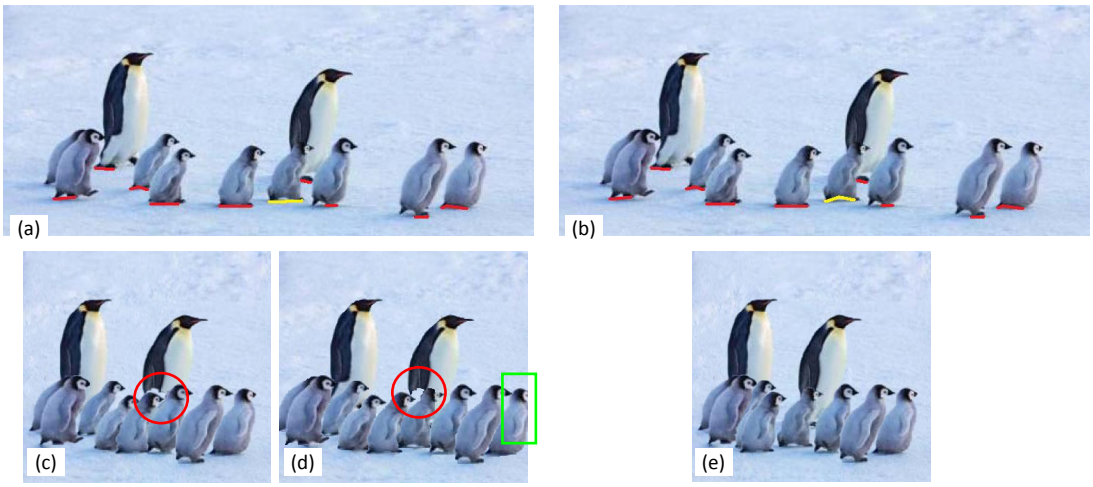


Figure 5.9. (a)–(b) Input images with object handles, except that the yellow handle is non-rigid in (a) and rigid in (b). (c)–(d) Results of tearable warping (using handles in (a)) and scene carving, where both show hole artifacts (in red circles). (e) Results of tearable warping (using handles in (b)), where the yellow rigid handle has prevented the hole from showing up in the retargeted image. The green rectangle highlights unpleasant cropping of object in scene carving.

preserved. The dancers example in Figure 5.8 is one such cases (the handles are on the floor).

A potential problem with both scene carving and tearable warping is that, at extreme retargeting ratio, unpleasant hole artifacts can happen due to the dis-occlusion of objects, as shown in Figure 5.9(c) and 5.9(d). Despite the use of an energy function to minimize holes in scene carving, this hole problem is still unavoidable in extreme retargeting. For tearable image warping, this problem can be avoided by creatively specifying the object handles. For example, by changing the handle of the penguin blocking the occluded penguin to a rigid handle, no more hole (or inpainting artifact) is visible in the retargeted image as demonstrated in Figure 5.9(e).

Another creative use of handles in tearable warping is to select only part of a physical object for protection and use the handle to ensure that the selected part of the object is later combined seamlessly with the non-selected part in the retargeted image. This can be particularly useful for images where the entire object fills most of the image frame like the example of the Eiffel Tower in Figure 5.10. Scene carving does not have this flexibility as it does not guarantee the semantic connectedness between different parts of the object.

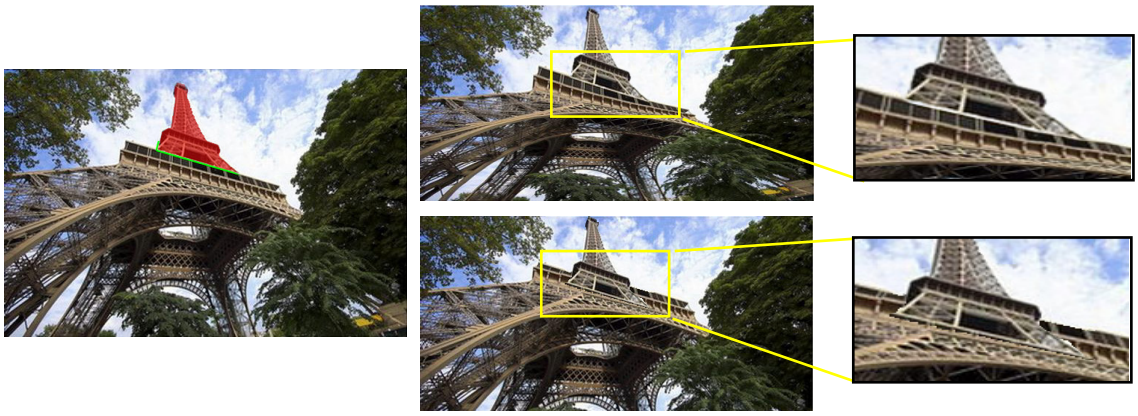


Figure 5.10. Creative use of object handles. (Top row) Input image with only the top part of Eiffel Tower selected for protection (handle is shown in green), (middle row) result of tearable warping, and (bottom row) result of scene carving.

5.3 Image Recomposition

We provide two operation modes for image recomposition; semi-automatic and interactive. In semi-automatic image recompositon, our algorithm embeds a set of aesthetics-distance energy based on selected photographic rules to drive the warping process to automatically modify the image compositon. For interactive

image recomposition, users can move selected objects and background interactively to recompose the image.

5.3.1 Semi-automatic Image Recomposition

In Section 2.1.1, we have identified six prominent aesthetics elements as illustrated by Figure 2.1. Of these six aesthetics elements, only three aesthetics elements – *subject dominance*, *balance* and *geometric elements* can be achieved by changing the image composition. An ideal aesthetics-driven image recomposition should aim to modify the image composition to make it adhere to photographic rules intended to enhance all these three aesthetics elements. However, unlike composing an image in the real, three-dimensional world, not all rules can be used in digital image recomposition since image modification is limited to the content within the image frame. For example, rules such as framing, leading lines and S-curve cannot be created by recomposition if these features do not exist in the original image. As depicted in Table 5.1, we singled out five photographic rules that can be used to enhance image composition.

In this work, we focus on four of these photographic rules namely, *simplicity, rule-of-thirds, visual balance*, and *fill the frame*. Enhancement of image perspective is excluded in the current work as it requires special consideration and applies to a smaller category of images, particularly cityscape or indoor scene. Nevertheless, we ensure that the proposed approach can be extended to support this aesthetics feature. To maximize image aesthetics, we aim to produce an output image that adheres more closely to the selected photographic rules. In other words, we attempt to reduce the distance between the composition of the output image and each photographic rule. Therefore, we formulate a set of aesthetics measures based on

Table 5.1. Photographic rules and corresponding image operations required.

Aesthetics measures		**Image operation required for recomposition**
Aesthetics Elements	**Photographic composition rules**	
Subject Dominance	Fill the Frame	*Change size of photo subject*
	Simplicity	*Change background of the photo subject*
	Framing	×
	Leading Lines	×
Balance	Rule-of-thirds	*Move photo subjects or horizon*
	Visual balance	*Move photo subjects*
Geometrical Elements	Lines / S-Curves	×
	Image Perspective	*Change the vanishing point*

the selected photographic rules in the form of aesthetics-distance energy, such that it can be minimized during the warping process to guide image recomposition to produce the maximally-aesthetics version of the input image as the result.

5.3.1.1 Aesthetics-Distance Energy

The set of aesthetics-distance energy used for automatic recomposition consists of subject dominance energy, rule-of-thirds energy, visual balance and size energy. Notably, an important contribution of this dissertation is in formulating the novel subject dominance energy to guide the image recompositon to effectively enhance the visual dominance of a photo subject

5.3.1.1.1 Subject Dominance Energy

Making the subject dominant. The core purpose of simplicity rule is to enhance subject dominance of the main photo subject and thus, image simplicity and subject dominance are highly correlated. Therefore, we measure simplicity in the form of subject dominance energy by adopting a simplified adaptation of the Itti-Koch biological-inspired visual saliency model (Itti et al. 1998). We treat the photo subject as the center and its immediate background as the surround, and measure subject dominance by computing the center-surround differences of two low-level features—luminance and color—between the subject and its background. The Lab color space is ideal since the L*, a* and b* components represent the luminance, the color position between red and green (R-G) and the color position between yellow and blue (Y-B) respectively. A highly dominant subject should exhibit high center-surround contrast in one or more of these intensity and color features.

To ensure efficient integration with the triangle-based warping method, we use a triangle mesh to represent the background image. Ideally, center-surround differences for the L*, a* and b* features should be measured between neighboring triangles connecting the subject and its immediate background. However, as the subject is not warped with the background and the set of background triangles neighboring the subject can change dynamically during optimization, it is not efficient to represent the object as a triangle mesh and to recompute its new neighboring background triangles for each iteration of the optimization. Therefore, we compute the center features by finding the average L*, a* and b* features for the whole subject. The set of surround features consists of the L*, a* and b* features for each triangle in the background within an expanded bounding box of the object. Using this center-surround contrast measure to guide tearable image warping, the

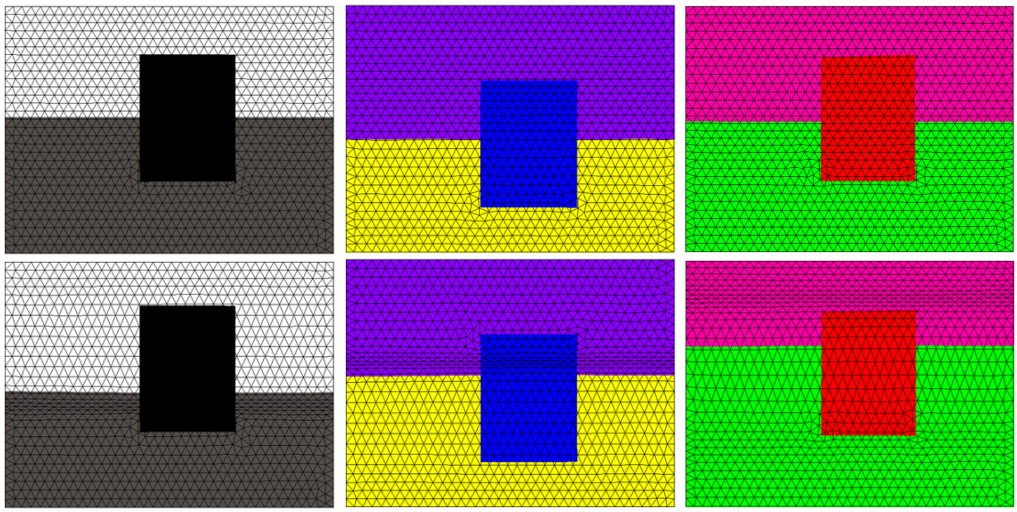


Figure 5.11. (Top) Synthetic input images that possess luminance and color contrast. (Bottom) Results of our algorithm show increased visual dominance of the rectangle.

algorithm makes the photo subject more dominant by enlarging background triangles that exhibit high contrast with the photo subject while compressing background triangles with low center-surround contrast. The synthetic examples in Figure 5.11 illustrate the effectiveness of our approach to increase the contrast of the rectangular subject. Results of our approach on natural images (e.g. Figure 5.12) have shown this approximation to be still effective.

Subject dominance energy: To maximize the luminance and color contrast, we minimize the *subject dominance energy*. We compute this energy only for the main photo subject and represent the luminance and color features in the Lab color space. Given the set of background triangles, T_o for the main photo subject t*o*, we define the *subject dominance energy* as

$$E_D = E_L + E_C, \tag{5.6}$$

where E_L is the *luminance contrast energy*:

$$E_L = \sum_{t \in T_o} s_t(|L'_t - L_o| - \psi_L)\,, \tag{5.7}$$

and E_C is the *color contrast energy*:

$$E_C = \sum_{t \in T_o} s_t\left(\sqrt{(a'_t - a_o)^2 + (a'_t - b_o)^2} - \psi_{ab}\right). \tag{5.8}$$

L'_t, a'_t and b'_t are the average values of L*, a* and b* of each triangle $t \in T_o$ in the output mesh and L_o, a_o and b_o are the average values for the main photo subject $o \in O$. s_t is the area scaling applied to each original triangle $t \in T_O$. It is set to $s_t = s_t^y$ if scaling of triangles is allowed only in the vertical dimension, or $s_t = s_t^x s_t^y$ if in both vertical and horizontal dimensions. In our implementation, we use only vertical scaling. Parameters ψ_L and ψ_{ab} are indicators of the target level of dominance, where

$$\psi_L = \left(\frac{1}{|T_o|}\sum_{t \in T_o}|L_t - L_o|\right) + \mu\,, \tag{5.9}$$

$$\psi_{ab} = \left(\frac{1}{|T_o|}\sum_{t \in T_o}\sqrt{(a_t - a_o)^2 + (b_t - b_o)^2}\right) + \mu\,, \tag{5.10}$$

and L_t, a_t and b_t are the average values of L*, a* and b* of each triangle $t \in T_o$ in the input mesh. μ is a parameter related to the target level of dominance and its value

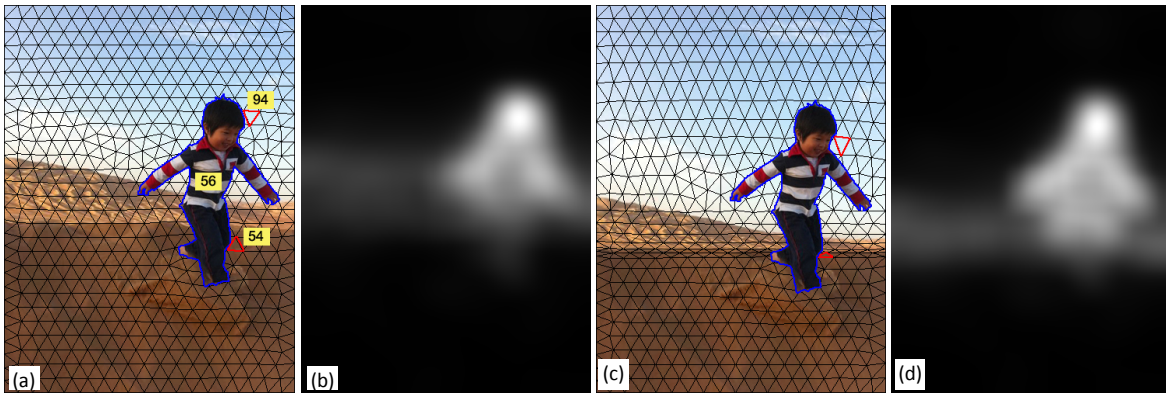


Figure 5.12. (a) Input image with triangle mesh. Yellow rectangles show the average luminance (L*) values in the subject and the red triangles. (b) Output image. Triangles have been expanded and compressed accordingly. (c& d) Saliency maps from input and output images respectively.

changes across images, depending on two features—background contrast and subject-background contrast. To obtain the μ value, we create a set of synthetic images with fixed background contrast but varying subject-background contrast. For each image, we then find the μ value that produces result with the maximum subject-background difference, as illustrated in Figure 5.13. By plotting the maximum μ value against the standard deviation of subject-background difference and fitting a graph to this plot, we obtain a formula for μ. We repeat this process for a few sets of images with each image set having a different background contrast.

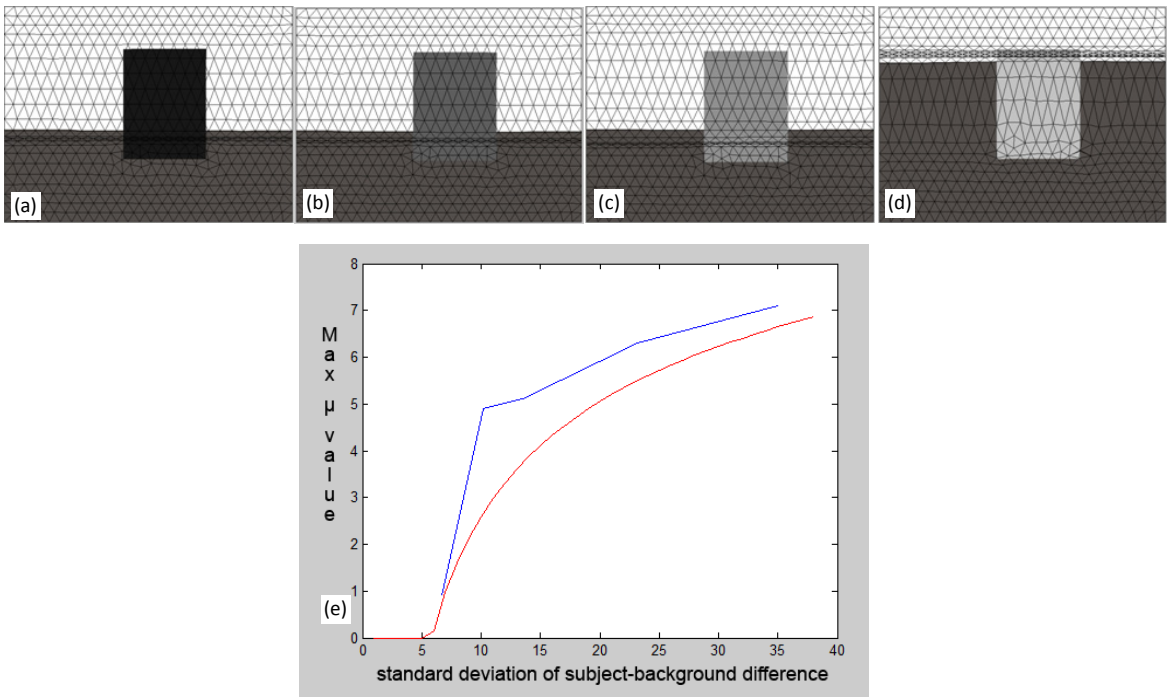


Figure 5.13. (a–d) Max $\boldsymbol{\mu}$ value for each image with the same background contrast but varying subject-background contrast is found by compressing each image to the maximum. (e) (Blue) The plot of max $\boldsymbol{\mu}$ value (y-axis) against the standard deviation (x-axis) of subject-background contrast. (Red) We obtain a formula for $\boldsymbol{\mu}$ value by fitting a logarithm function to the plot.

We then find a graph that fits reasonably well to each plot and obtain the following adaptive μ value:

$$\mu = \frac{S(F)B(F)}{\tau} - \frac{B(F)^2}{18*S(F)}, \tag{5.11}$$

where F is the feature (L* or ab*), $S(F)$ is the subject-background contrast, given by the standard deviation of the center-surround feature differences, and $B(F)$ is the background contrast for feature F, given by the standard deviation of the background triangle features. The value of τ is dependent on features. We use $\tau = 9$ and $\tau = 6$ for luminance and color features respectively.

5.3.1.1.2 Rule-of-thirds Energy

The *rule-of-thirds energy* comprises of two components; the *power-point energy* and the *horizon energy*. The *power point energy*, E_G pulls objects toward one of the four power points and is defined as,

$$E_G = \sum_{o \in O} A_o D_P(o), \tag{5.12}$$

where $D_P(o)$ is the minimum distance from the object weighted centroid to the four power points and A_o is the normalized object size. If the optional face direction constraint is not specified, all four power points will be considered in Eq (5.12). Otherwise, the power points being used in Eq (5.12) is selected based on the face direction such that sufficient space is left in front of the subject.

The *horizon energy*, E_H minimizes the distance between the new horizon, $\widehat{H}$ and the nearer horizontal power line, PH_j, and is defined as

$$E_H = \min_{j=\{1,2\}} (\left|\widehat{H} - PH_j\right|). \tag{5.13}$$

If horizon does not exist, then $E_H = 0$.

5.3.1.1.3 Visual Balance Energy

The *visual balance energy*, E_v is formulated as

$$E_v = \sum_{o \in O} A_o \left\| \hat{C}(I') - \hat{C}(o_i) \right\|_1 , \quad (5.14)$$

where $\hat{C}(I')$ is the image center and $\hat{C}(o)$ is the weighted centroid of object o. For images with only one object, $E_v = 0$.

5.3.1.1.4 Size Energy

The *fill the frame* rules aims to enhance subject dominance by filling the image frame with the photo subject as much as possible. However, we observe that this rule is catered more for portrait and macro images. For other type of images such as landscape, the size of the photo subject is only a fraction of the image frame. Based on an experiment conducted by Liu et al. (2010), the size of region of interest (ROI) in professional photographs is non-uniformly distributed. Figure 5.14 illustrates the histogram of the sizes of automatically detected ROI in a database of more than 200 professional images. The histogram distribution has three dominant peaks. This study infers that the aesthetically pleasing sizes of the main photo subject are mostly distributed around the set of sizes, S = {0:1, 0:56, 82}, that correspond to small, medium and large regions respectively. Therefore, we formulate the *size energy*, E_S to gravitate the new size of the main photo subject, $\widehat{S_o}$ towards the nearest aesthetically pleasing size, S_j.

$$E_S = \min_{j=\{1,2,3\}} \left(\left| \widehat{S_o} - S_j \right| \right). \quad (5.15)$$

To preserve image semantics, we maintain the relative size difference among all subjects by changing the size of secondary subjects according to the new size of the

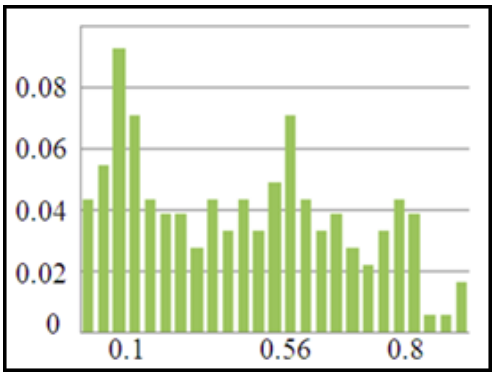


Figure 5.14. Aesthetically pleasing sizes of professional photos are distributed around 3 peak sizes; 0:1, 0:56, and 0.82 (Liu et al., 2010).

main photo subject. In addition, we constrain the maximum scale of the photo subject to 1.4 of its original size to avoid serious loss of image resolution.

5.3.1.1.5 Total Aesthetics-Distance Energy

For aesthetics optimization, we aim to minimize the total aesthetics-distance energy defined as

$$E_A = w_D E_D + w_G E_G + w_H E_H + w_V E_V + w_S E_S\,, \tag{5.16}$$

where w_D, w_G, w_H, w_v and w_s are the weights for each aesthetics-distance energy defined in the previous sub-sections.

5.3.1.2 Recomposition-specific Constraints

Object boundary constraint. We first compute an axis-aligned bounding box around the object. During optimization, we compute the new position of the bounding box based on the new position of the object handle. To enforce this

constraint, we disallow any part of the object's bounding box to move outside the target image region.

Foldover constraint. To avoid triangle foldover problem, we limit the scale factor of each triangle mesh to 0.15.

5.3.1.3 Total Energy

To perform automatic image recomposition, we warp the image by minimizing the total energy, E that consists of the scale transformation error, E_w, the smoothness error, E_s and aesthetics-distance energy E_A,

$$E = \alpha E_w + \beta E_s + \gamma E_A. \tag{5.17}$$

where α, β, and γ are the weights for the respective energy component.

5.3.1.4 Implementation

User Input and Interaction. Apart from object segments and their respective object handles, for semi-automatic recomposition, users can provide two optional input; horizon and face direction. The horizon in an image can be automatically detected using (Hoeim et al. 2007) and easily modified by users by dragging the detected horizon line. The face direction can be specified with a single click.

Optimization details. We use the CVX Matlab toolbox (Grant et al.) to find the solution to the convex quadratic function defined in Eq. (5.17). The weights for the total energy, α, β and γ are set to 1, 0.5 and 1. The weights for the aesthetics-distance energy, w_D , w_G , w_H , w_V and w_S are set to 0.8, 0.5, 0.1, 0.5 and 0.8 respectively. The handle shape and image boundary constraints are set as hard

constraint. To prevent objects from being cropped off, the object boundary is set as an inequality constraint.

5.3.1.5 Experimental Results

We tested our algorithm on a set of test images selected from our personal photo collections, from Flickr.com, and from the result set of the crop-retarget approach (Lit et al. 2010). Results were generated on a laptop with Intel Core2 Duo CPU 2.53GHz. Excluding time for inpainting, it took about 10 seconds to produce a recomposed image with resolution 800 x 600.

5.3.1.5.1 Results

Overall, by capacitating the change of subject-background relationship around the tearable boundary segments, we observe that our results showcase significant change in image composition without violation of spatial semantics. Oftentimes, the resulting images look as if they have gone through a natural change of viewpoint from the input image as illustrated in our results in Figure 5.15. Notably, testing shows that our results are not very sensitive to the change of parameter values. We compare our results with the results of crop-retarget (Liu et al. 2010), which, to our knowledge is the best automatic recomposition operator among the state-of-art recomposition methods.

Our results illustrate the effectiveness of our algorithm in modifying the composition of an input image to adhere to specific photographic rules. Comparing the saliency maps of our results to the corresponding saliency maps of the input image in Figure 5.15, we can see that the background of the resulting image is less distracting, making the photo **subject more dominant**. In contrast, due to the static

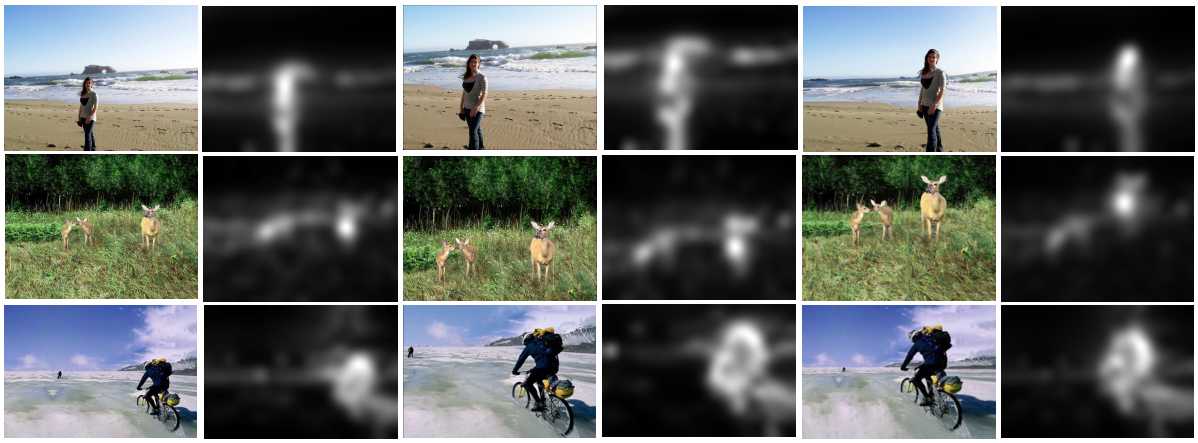


Figure 5.15. (column 1 and 2) Input images and the saliency maps (column 3 and 4) Results of crop-retarget (Liu et al. 2010) and the saliency maps (column 5 and 6) Results of tearable image warping, and the saliency maps.

subject-background relationship nature of crop-retarget operator (Liu at al. 2010), saliency maps of results produced by crop-retarget operator show little change in background saliency of the resulting images. Comparatively, our approach using tearable image warping has more potential to produce significant change in image recomposition.

Adherence to **rule-or-thirds** is seen in almost all our results in Figure 5.15 and 5.16, whereby photo subjects are moved nearer to one of the four power points. Interestingly, we observed that the face direction constraint that we apply to rule-of thirds energy directs our algorithm to generate more pleasing results in some images such as the beach and penguin examples in Figure 5.15 and Figure 5.16 respectively. Without any face direction constraints, as in the results of the crop-retarget method, there is little space left in the facing direction of the lady and the penguin, invoking an unpleasant feeling in viewers. Apart from photo subject, we find that the horizon line in many of our results has been repositioned near one of the power lines.

Figure 5.16. More results. (row 1) Input images. (row 2) Results of crop-retarget (Liu et al. 2010) (row 3) Results of tearable image warping.

Improved **visual balance** is evident in the swans and elephant examples in Figure 5.16. In addition, unlike the crop-retarget operator in which cropping is performed to change the size of the photo subject, our approach successfully alters the **size** of photo subjects without sacrificing the global context. The cycling example in Figure 5.15 and the boat and bird examples in Figure 5.17 clearly illustrate the strength of our size change approach to preserve the global context. On the other hand, for images with lots of redundant background such as the lady and the windmill examples in Figure 5.17, cropping based method such as crop-retarget has the edge to produce more pleasing results. More results of our approach are shown in Figure 5.18.

5.3.1.5.2 User Study

For an objective evaluation of the effectiveness of the tearable image warping approach in automatic recomposition, we conducted three online human subject experiments; one experiments to validate its effectiveness in enhancing subject dominance and two experiments to validate its ability to enhance image aesthetics.

5.3.1.5.2.1 Validation of Subject Dominance

The objective of the first experiment, Experiment 1 is to validate the effectiveness of the subject dominance energy in driving tearable image warping to enhance visual dominance of the photo subject. For this experiment, we obtained a set of the 30 test images described in Section 5.3.1.5 and apply the recomposition algorithm with a single aesthetics-distance energy – the subject dominance energy, to produce a set

Figure 5.17. More results. (row 1) Input images. (row 2) Results of crop-retarget (Liu et al. 2010) (row 3) Results of tearable image warping.

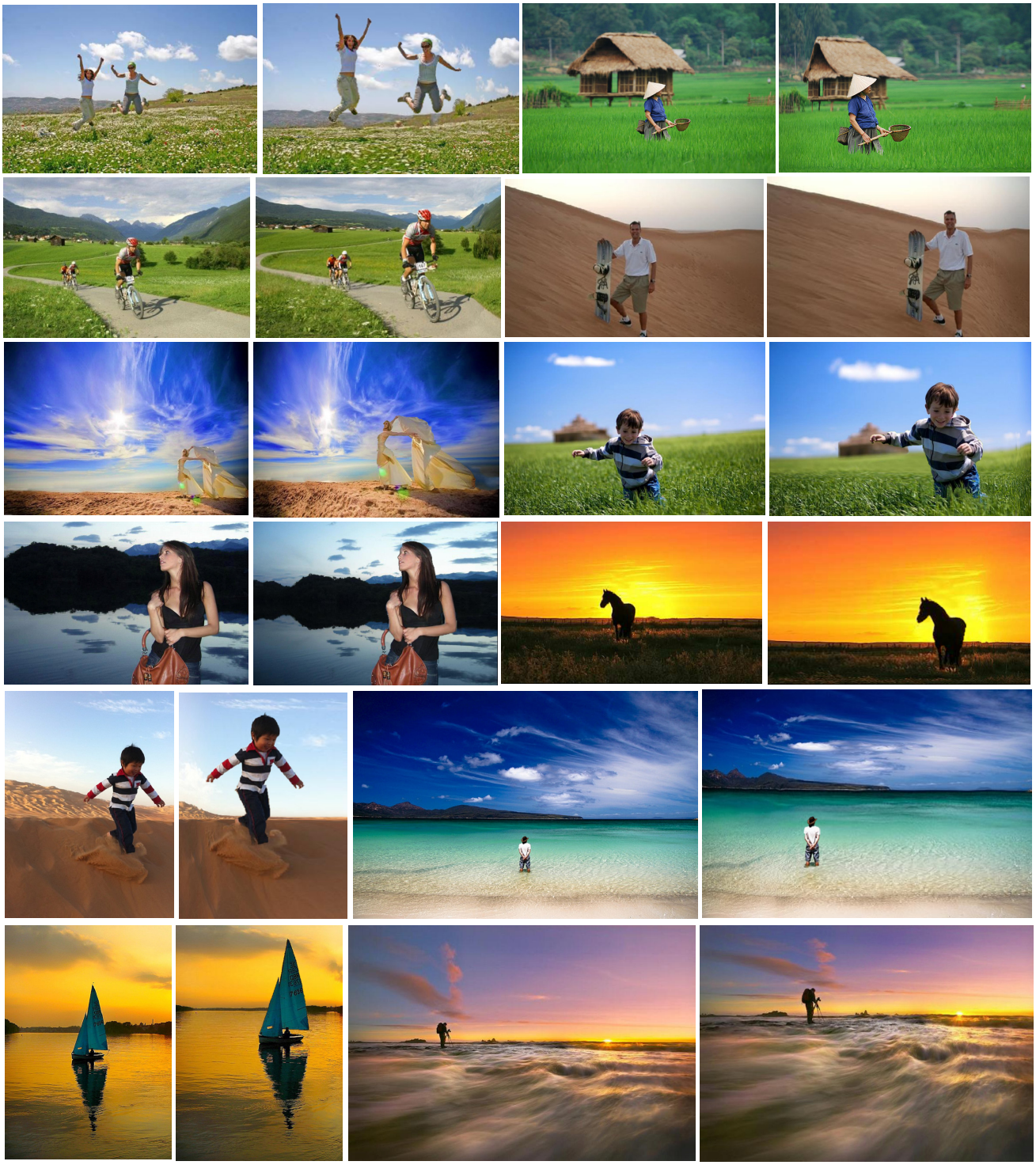


Figure 5.18. More results. (left) Input images. (right) Results of tearable image warping.

of corresponding recomposed images. Figure 5.19 shows some image pairs used in the experiment. We let each participant compare these 30 pairs of images. Each time, an input image and our result were shown side-by-side, with left-right positions randomly chosen. The participants were instructed to choose one where the photo subject stands out more against the background. The experiment had 40 participants, consisting of males and females aged between 22 and 46.

Results of Experiment 1: The outcome is, for 83% of the image pairs, our results were chosen. This result demonstrates the effectiveness of our recomposition approach to enhance subject dominance, which potentially leads to improved image aesthetics.

5.3.1.5.2.2 Validation of Aesthetics Enhancement

We conducted two online user experiments, Experiment 2 and Experiment 3 to

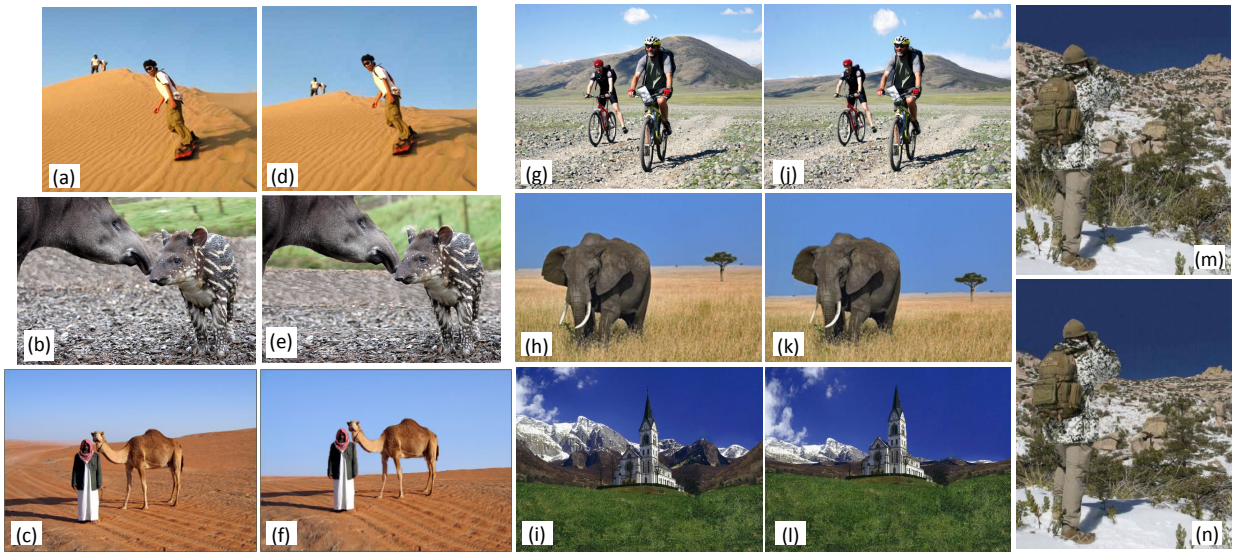


Figure 5.19. (a–c, g–i, m) Input images. (d–f, j–l, n) Results from our algorithm with only subject dominance energy.

study the effectiveness of our method in enhancing the aesthetics of the images. In Experiment 2, we compare our results with the original input images. In Experiment 3, we compare our results against crop-retarget approach (Liu et al. 2010), the leading state-of-art automatic recomposition approach. The images we used in the user experiments are a subset of the test images mentioned in Section 5. 3.1.5. We use 50 images for Experiment 2 and 35 images for Experiment 3. There are 25 overlapping images in these two image sets and thus, the combined results from these two experiments can be used to infer the relative aesthetics preferences among the original image, crop-retarget method and our approach. In both experiments, each image pair is shown side-by-side. The screen positions (left or right) of the images in each image pair are chosen at random. For each image pair, subjects are asked to choose the image that they think is more aesthetically pleasing from the following set of choices; {"left image is much better", "left image is slightly better", "both images are similar", "right image is slightly better" and "right image is much better"}. A total of 50 test subjects aged between 18 and 51 participated in each experiment.

Results of Experiment 2: Results for this experiment is encouraging. The graph in Figure 5.20(a) shows that, for 66% of the image pairs, majority of the test subjects chose the recomposed image using our approach compared to only 30% for the competing input images. For further analysis, we assign a score for each user selection; 2 and 1 if a recomposed image is chosen as "Much Better" and "Slightly Better" respectively, 0 if both images are deemed as "Similar", and -2 and -1 if an input image is chosen as "Much Better" and "Slightly Better" respectively. We then sum up the scores from all test subjects for each image and plot the area graph shown in Figure 5.20(b). The area under the graph for positive score difference

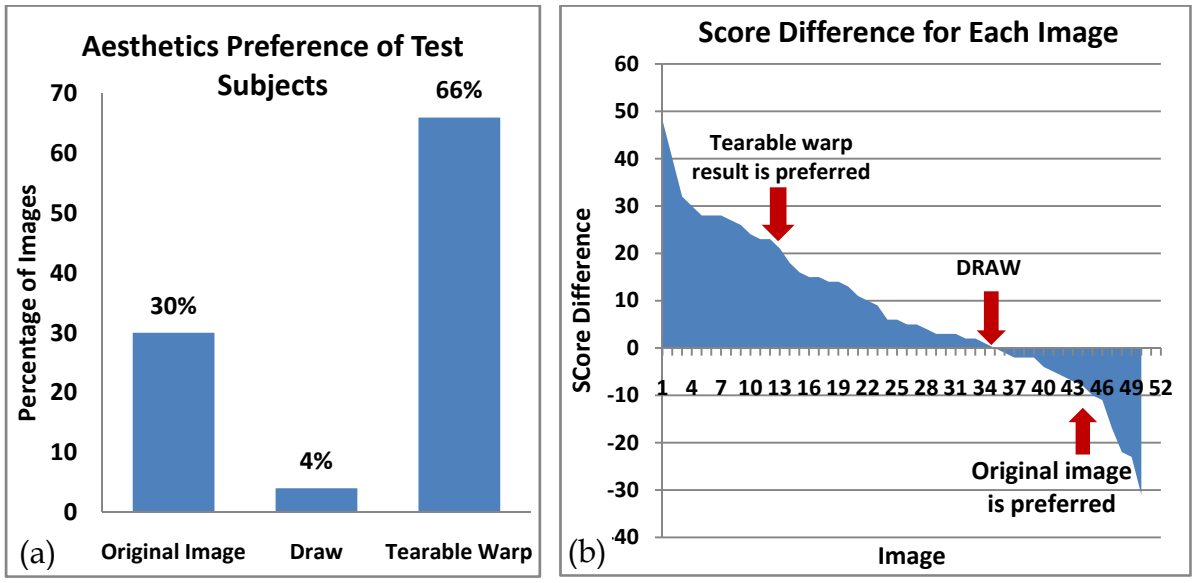


Figure 5.20. Tearable image warping VS Original image (a) Image preferred by majority of the subjects. (b) Score difference between tearable image warping and crop-retarget for each image.

represents the sum of scores difference when results of tearable image warping are the preferred choice. Vice versa, the area under the graph for negative score difference depicts the sum of scores difference when input images are preferred. It is obvious that the score difference is significantly larger for cases when the recomposed image is preferred. If we consider the images with score difference less than 10 as noise where the subjects don't have significant preferences, test subjects have a significant preference for the input images only for 10% of the image pairs. These images are shown in Figure 5.21. Upon further analysis, we discovered the reasons for the lack of preference for these few results of our approach. Firstly, over-compression due to large displacement of photo subjects produced unpleasant, minor artifacts in the area highlighted by the red rectangles in images shown in Figure 5.21. Secondly, the non-linear warping applied to the background could sometime cause some parts of the background lose their interestingness such as the

lake formation and footsteps highlighted by the yellow scribbles in the examples in Figure 5.21. Furthermore, placing the photo subject too near the border also invokes an unpleasant feeling, as in the man by the lake example in Figure 5.21.

Despite the infrequent cases of reduced aesthetics mentioned above, oftentimes, maximizing the aesthetics-distance energy successfully leads to improved image aesthetics in the resulting images. Notably, test subjects have an clear preference for images that has significantly improved visual dominance due to change of subject-background relationship, such as the images presented in Figure 5.15 and jumping girls and lady by the lake examples in Figure 5.18. In addition, we observed that another aesthetics pulling factor is the change in size of the photo subject which also effect in increased visual saliency, such as the elephant and swans examples in Figure 5.16 and the boat, bird and sunset beach examples in Figure 5.17.

Figure 5.21. Limitation of our recomposition approach. (top) Input images. (bottom) Results of our approach. Red rectangles highlight the over-compressed area resulting in unpleasant visual effect. Red line points out that the object is too near the highlighted border. Yellow scribbles highlight the interesting background which is not preserved in the recomposed image.

Nevertheless, balance elements which include rule-of-thirds and visual balance, undoubtedly also contribute to improving the overall composition of the resulting images, leading to improved aesthetics.

Results of Experiment 3: From the graph in Figure 5.22(a), we observe that our approach performs slightly better than the state-of-art crop-retarget (Liu et al. 2010) method, with a score of preference of 54% for results of our approach compared to 43% for results of crop-retarget. The score difference for both the crop-retarget approach and our approach is not significantly different. This result is not surprising as both our approach and crop-retarget approach have their own strength and potentially would work better for different category of images.

Examining the resulting images where test subjects has significant preference for either the tearable warp or crop-retarget approaches provide some valuable insights into the type of images that are well-suited for each approach. Features that set the

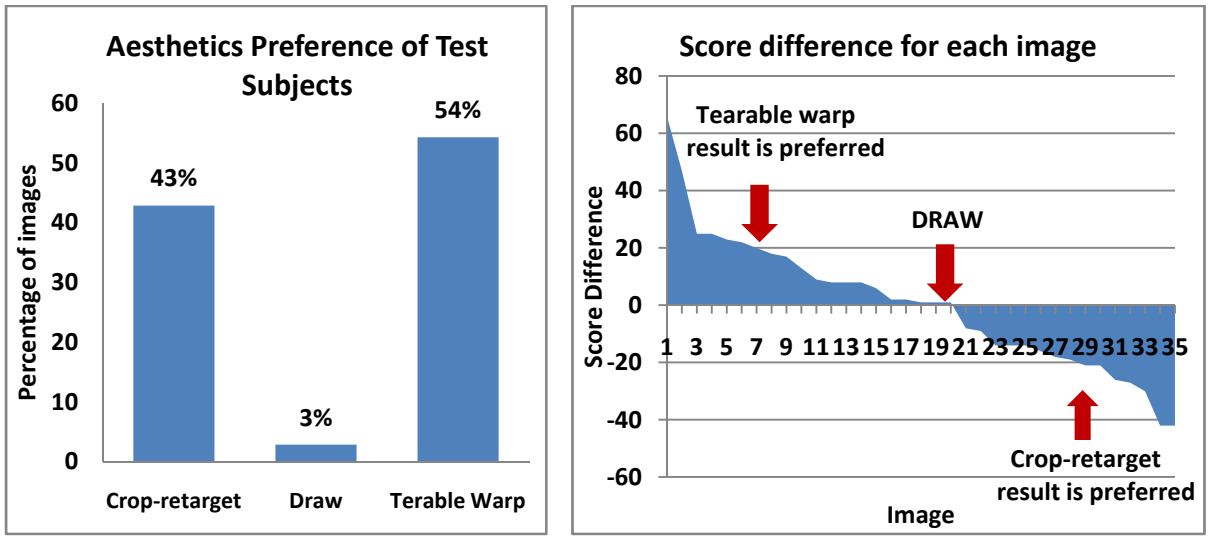


Figure 5.22. Results of Experiment 3 – Tearable image warping VS Crop-retarget (left) Image preferred by majority of the subjects. (right) Score difference so each image.

tearable warp approach apart from crop-retarget method are its ability to preserve the global context, to protect photo subjects and to enhance visual dominance of photo subjects. Inevitably, these are the distinctive features found in results of tearable warp that obtain significant preference from test subjects. Many results of tearable warp like the boat and birds examples in Figure 5.17, won over crop-retarget approach due to the preservation of global context and interesting background that was cropped off by the crop-retarget approach. Interestingly, the ability to preserve interesting background is so important that test subjects would not mind some minor distortion. For example, subjects have a clear preference for the results of our approach for the squirrel and girl on wheelchair images in Figure 5.23 despite the distorted line and over-compressed background.

In addition, our recomposition approach scores well in cases where visual dominance can be enhanced without sacrificing the interestingness of the

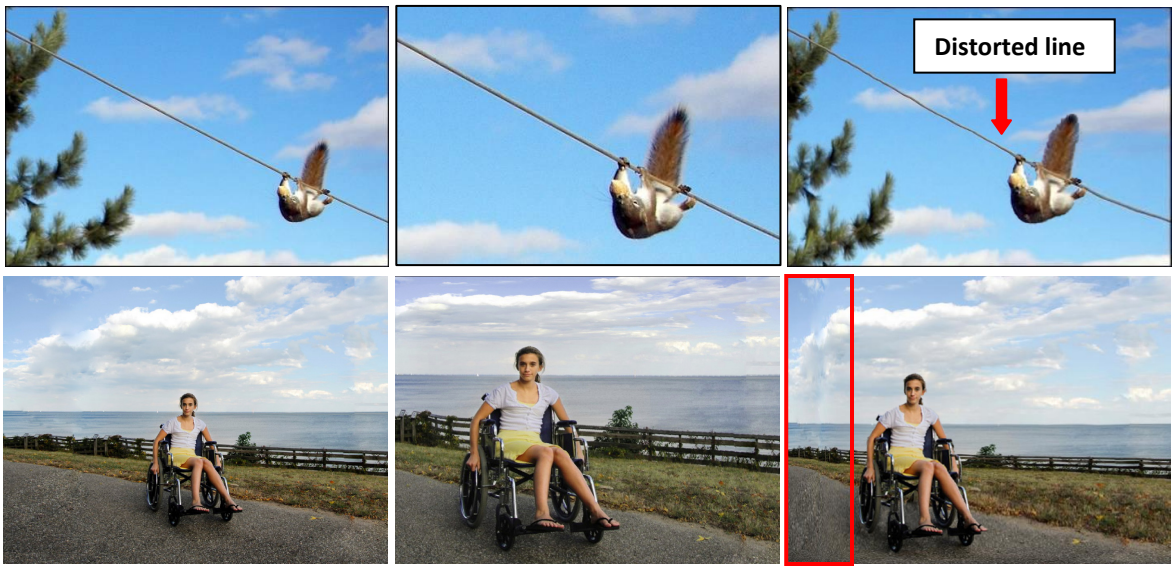


Figure 5.23. Global context VS minor distortion. (left) Input images. (middle) Results of crop-retarget. (right) Results of our approach. Red rectangle highlights minor artifacts caused by over-compression.

background as illustrated by the examples in Figure 5.15. Having said that, it is important to note that increasing subject dominance at the expense of compressing interesting background can potentially lead to reduced aesthetics preferences as in the couple and castle examples in Figure 5.16 and Figure 5.17 respectively. Another advantage of our approach is the guaranteed object protection that can avoid object distortion problem that potentially occurs in crop-retarget, as seen in the elephant example in Figure 5.16 and the girl on wheelchair example in Figure 5.23.

On the other hand, the crop-retarget works better for images where there exists uninteresting background that can be cropped away to bring more focus to the subject as illustrated by the sunset and windmill examples in Figure 5.17. Comparatively, crop retarget is less prone to distortion as the cropping component of this hybrid approach reduced artifacts induced by warping as illustrated by the girl on wheelchair example in Figure 5.23.

In summary, tearable warp approach is applicable for images in which preservation of the global context or enhancement of visual dominance is pertinent to ensure enhanced aesthetics experience. In contrary, crop-retarget approach would be more befitting for images with significant redundant or uninteresting background. Notably, adherence to any photographic rules, particularly subject dominance should not be made at the expense of sacrificing interesting background elements that may contribute to the overall aesthetics experience. This finding sets the foundation for our future work to further improve the robustness and effectiveness of our recomposition approach.

Combined results of Experiment 2 and Experiment 3: From the 25 overlapping images from Experiment 1 and 2, we inferred and compared the aesthetics preference of the test subjects. Interestingly, we find that for 96% of the images,

subjects chose either results of tearable image warping or crop-retarget as shown in the graph in Figure 5.24. Of this total percentage, 50% of the results of tearable warp are preferred compared to 46% for results of crop-retarget. This result is consistent with the results of Experiment 2 and reinforces that tearable image warping is indeed a good complementary approach for crop-retarget and both approaches can be targeted for editing different type of images.

5.3.2 Interactive Image Recomposition

For interactive image recomposition, users can perform two operations; **object relocation** and **background warping**. To relocate an object, users select the object and move it to the desired location. To warp the background, users can drag any part of the background and move it in the desired direction. This extra background warping feature allows users to have more control over the composition of the

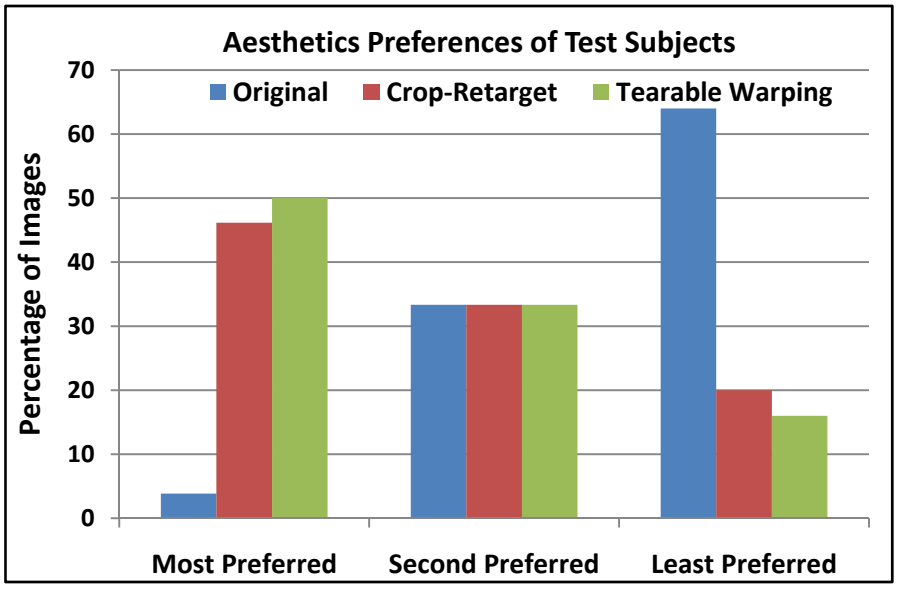


Figure 5.24. Combined results from Experiment 2 and Experiment 3.

output image. For example, users can move the horizon to a desired location by simply dragging the background to a new position as shown in the example in Figure 5.25.

5.3.2.1 Interactive Recomposition-specific Constraints

Object Relocation. For object relocation in recomposition, the user can interactively move any selected object, and even move it to occlude other objects. Due to the interactive nature, object boundary and non-overlap constraints are not necessary. The only required constraint is the handle positional constraint. We set the position of the selected object based on the user's mouse movement and fix the other objects at their current positions.

Handle positional constraint. Suppose each object's handle consists of n vertices, $v_1, v_2, \cdots, v_n$, in M, and they are being mapped to vertices $u_1, u_2, \cdots, u_n$ in M'. Let $d = (d_x, d_y)^{\mathrm{T}}$ be the translation computed from the mouse movement, the handle positional constraint for the selected object is

$$u_1 = v_1 + d, \tag{5.18}$$

and for every of the unselected objects, the constraint is

$$u_1 = v_1. \tag{5.19}$$

Interactive Background Warping. The method to enable interactive background warping is similar to that of relocating objects. In this case, the background is warped in the direction of the mouse movement while keeping all the object handles fixed at their current locations. More specifically, we first find the mesh vertex nearest to the mouse click position, and reposition the mesh vertex at the

Figure 5.25. Result of interactive background warping, where the horizon has been moved but the object's position fixed.

new mouse location using the same **handle positional constraints** as described above for object relocation.

5.3.2.2 Implementation

User Input and Interaction. In addition to object segments and their respective object handles, to interactively recompose a retargeted image, the user can click on any object and drag it within the retargeted image. The user can also modify the background by dragging any part of the background towards the targeted direction.

Optimization details. We formulate interactive recomposition as the problem of minimizing Eq. (5.3) with a set of equality constraints, for which the solution can be obtained in real-time by solving a sparse linear system. The weights for the total energy, α, and β are set to 1 and 0.5. The handle shape, image boundary and handle positional constraint constraints are all set as hard constraints.

5.3.2.3 Results and Discussion

We compare our results with aesthetics-driven, pure warping-based reecomposition approach with traditional warping based recomposition approach (Jin et al., 2010). From the bottom image in Figure 5.26, we can observe severe distortion above the lady's head in the results of pure warping. No similar distortion is detected in the corresponding result of tearable warping. In addition, the inability to allow changes to object-background relationship has limited the flexibility of traditional warping for image aesthetics enhancement. For example, the unpleasant effect of having the horizon cutting through the lady's neck in Figure 5.26 cannot be changed by traditional warping. With tearable warping, users can move objects or background to avoid merger or to effect a view change to make the subject more visually dominant. Figure 5.27 shows more examples of recomposition on retargeted images using tearable warping.

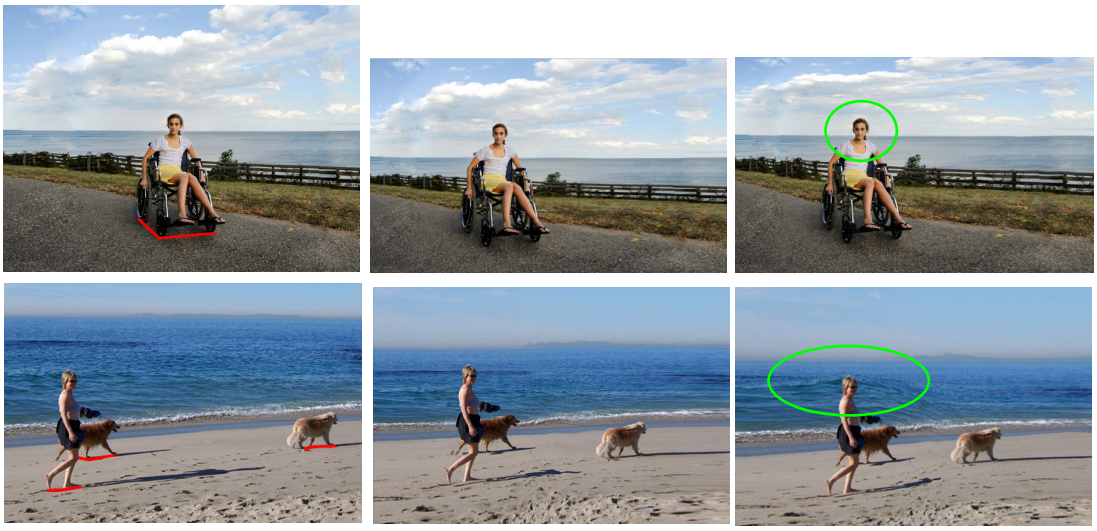


Figure 5.26. Results of interactive recomposition. (left) Input Images, (middle) results of tearable image warping, and (right) results of traditional image warping.

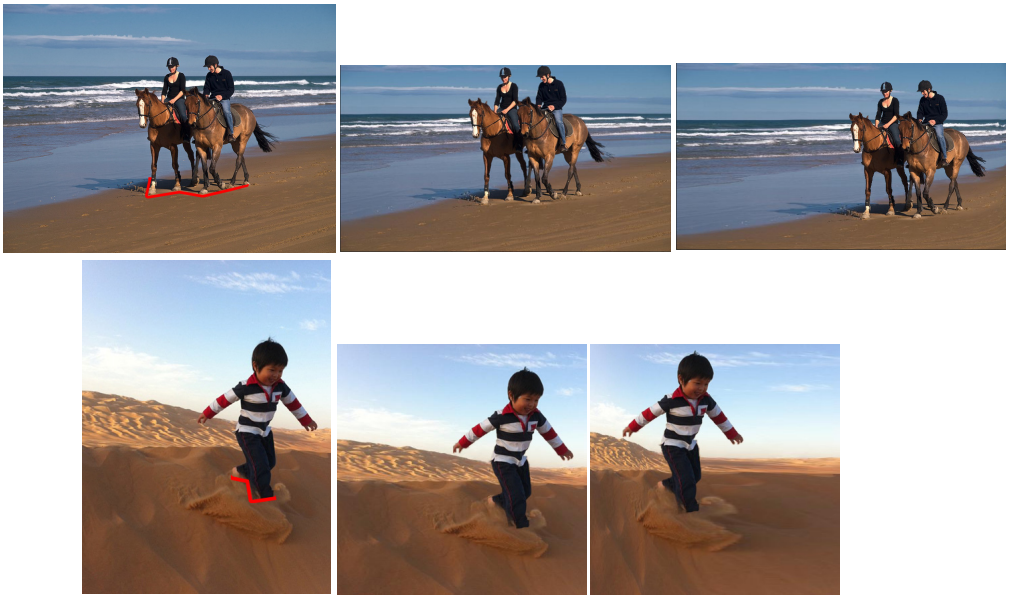


Figure 5.27. More results of interactive recomposition. (left) Input Images, (middle) retargeted images, and (right) recomposed results of retargeted images.

5.4 Limitation and Future Work

One drawback of the tearable image warping approach is that it requires inpainting, which is still an open problem in computer vision. Fortunately, this drawback is not that critical to tearable image warping, as compared to pure cut-and-paste approaches. We find that good inpainting is seldom required for tearable image warping, particularly for retargeting an image to a smaller size. Artifacts of inpainting often occur near the object handle or the object's center, and these areas are most likely still covered by the object in the retargeted or recomposed image. Furthermore, the holes are usually compressed with the background image in the retargeting process, making it even less likely to show up. As illustrated in Figure 5.28, although the inpainted background image is far from perfect, none of these

Figure 5.28. (Left) Retargeted and recomposed background layer. Red outline shows the inpainted hole, which is compressed in the retargeting process. (Right) Result with the object re-inserted. The red areas show that only small part of the inpainted hole is visible in the final image.

artifacts are actually exposed in the result. However, for retargeting an image to a larger size, artifacts of inpainting have higher potential to show up as illustrated in Figure 5.29. For cases where inpainting artifacts are visible, we allow users to interactively touch up the inpainting artifacts.

Another drawback of our method is that it requires segmentation and recompositing that is not always so easy. In our work, we used hard segmentation in which good segmentation can be difficult to achieve, particularly in cases where there is high feature similarity between the object boundary and background as illustrated by the top image in Figure 5.30(a). Thus, one limitation of this approach

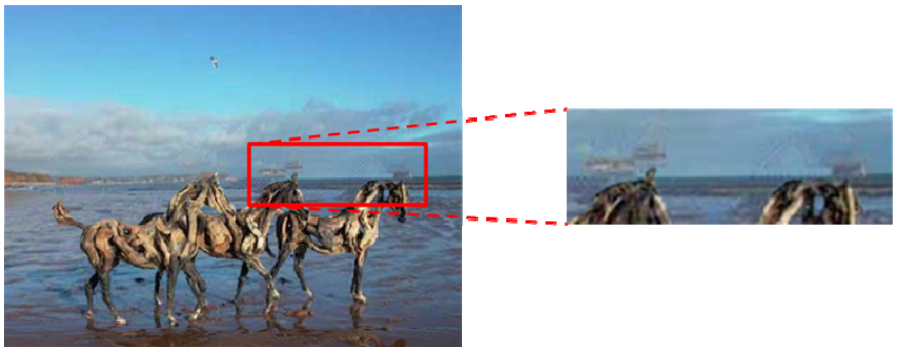


Figure 5.29. Artifacts of inpainting when retargeting an image to a larger size.

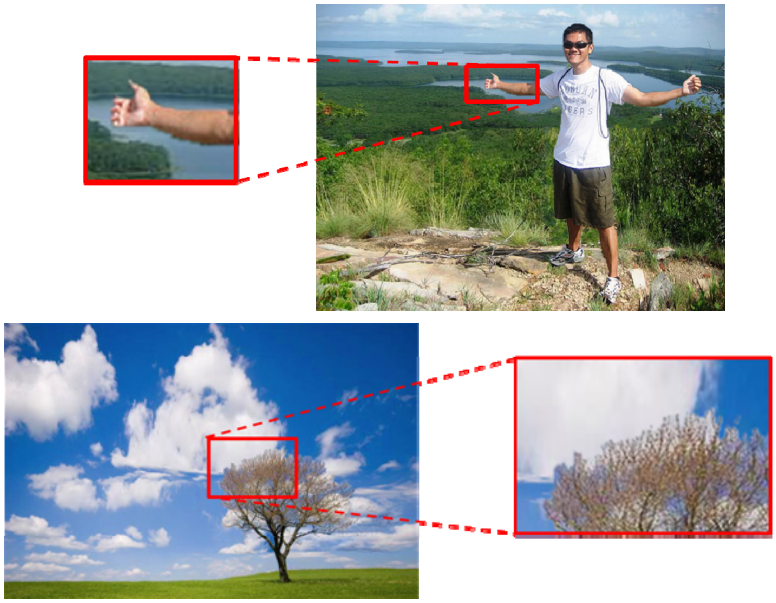


Figure 5.30. (Top) Artifact is inevitable in cases where there is high feature similarity between the object boundary and the background. **(Bottom)** Tearable image warping approach may not be suitable for images where the object is difficult to be segmented.

is that it may not work well for images when the foreground subject cannot be easily segmented. Figure 5.30(b) illustrates one such example. One possible solution to reduce the artifact at the object boundary is to employ digital matting to perform recompositing (Wang & Cohen 2007, Levin et al. 2008).

In our algorithm, we did not apply any feature preservation to preserve prominent lines and curves because we found that our results seldom have significant feature distortion, due to the fact that warping is spread out more uniformly throughout the image, particularly for image retargeting. However, for images with very complex structural details like the art room example in column one of Figure 5.7, distortion is hard to avoid in recomposition or extreme retargeting. In these cases, we can add an optional line preservation constraint to ensure that straight lines are not distorted.

For semi-automatic recomposition, we observed a few weaknesses of our approach as highlighted in Section 5.3.1.5.2.2. To solve these problems, we propose some minor refinement to enhance the robustness and effectiveness of our algorithm. First, to avoid the loss of interestingness in the background, our algorithm can detect and preserve the interesting background area by applying appropriate constraints to the relevant background area. In addition, we can enforce a stronger smoothness term which we believe can help to reduce both the loss of interestingness and to avoid unpleasant artifacts caused by over-compression. Lastly, to avoid undesirable effect due subjects being placed too near the image border, we can modify the object boundary constraint to ensure an offset from the image border in subject placement.

We obtained positive feedback on our interactive background warping approach and we foresee the potential of interactive background editing being adopted as a common image editing tool. However, the current implementation of our interactive background editing is quite trivial and may not be robust enough for advance background editing. There is much room for research in interactive background editing, particularly in capacitating more types of background editing and implementing trivial user interaction to support them.

5.5 Chapter Summary

We have introduced tearable image warping, a new approach that unifies image warping and cut-and-paste techniques, for content-aware image retargeting and recomposition. The key concept of tearable warping to allow an object to be partially detached from its original background makes several noteworthy

contributions. For image retargeting, it significantly reduces distortion inherent to traditional warping, particularly in cases of extreme retargeting. Besides, it can achieve better scene consistency by simultaneously protecting objects, ensuring correct depth order of objects and maintaining consistent semantics connectedness between objects and their environment. For image rcomposition, tearable image warping capacitates the change of object-background relationship, making it a powerful tool for significant image recomposition. In particular, tearable image warping supports our novel idea to implement a simplified center-surround contrast measure to guide the warping to enhance the visual dominance of the photo subjects in the recomposed images. In addition to the subject dominance energy, we applied a set of aesthetics-distance energy based on several photographic composition rules to guide the aesthetics enhancement in our image recomposition algorithm. Our results and user experiments have shown the effectiveness of our recomposition approach to enhance both visual dominance and image aesthetics.

Chapter 6

Conclusion and Future Research Direction

Beauty can be seen in all things, seeing and composing the beauty is what separates the snapshot from the photograph.

Matt Hardy

This thesis has presented several saliency-based approaches for both image aesthetics evaluation and enhancement. The three main contribution of this thesis include; saliency-based aesthetics evaluation models for aesthetics class and score prediction, the saliency retargeting algorithm for low-level image enhancement and the tearable image warping approach for extreme image retargeting and aesthetics-driven image recomposition. This chapter summarizes this thesis by giving a summary for each of these three works presented in the previous chapters and ends with proposed future research directions.

6.1 Summary

Motivated by the importance of subject dominance in influencing image aesthetics, the goal of this thesis is to utilize a saliency-based approach to effectively evaluate and enhance image aesthetics. In general, subject(s) are extracted from the image and features of the subject(s), particularly features denoting the dominance of the subject(s) are used in developing aesthetics evaluation model and aesthetics-driven image editing algorithms.

In Chapter 3, we presented the saliency-enhanced approach for aesthetics class and score prediction. By combining a set of subject-focused features with a set of prominent global features, we trained two aesthetics evaluation models; a classification model to discriminate professional photographs from snapshots and a regression model to infer an aesthetics score for a given image. Results show that our subject-focused approach significantly increases the accuracy of aesthetics class prediction compared to state of art approaches (Datta el al. 2006, Yan et al. 2006). Despite producing moderate correlation score, the aesthetics score prediction model successfully assists our saliency retargeting algorithm in maximizing the aesthetics of resulting images. This result demonstrates the effectiveness of the aesthetics score prediction model to infer relative aesthetics score of similar images and can be very useful in various applications including aesthetics-driven image editing and photo management systems.

Next, in Chapter 4, we introduced saliency retargeting, a novel low-level image enhancement approach aimed to enhance image aesthetics by redirecting viewers' attention to the important subjects of the scene. This approach applied non-uniform modification to three low-level image features; intensity, color and sharpness that

directly correspond to features used in biological plausible visual attention model (Itti et al. 1998). The aesthetics score prediction model presented in Chapter 3 was used to evaluate each enhanced image in a result image set and return the maximally-aesthetics version as the result. Empirical evaluations with human subjects demonstrate the effectiveness of our saliency retargeting algorithm in redirecting viewers' attention to important subjects, leading to enhanced image aesthetics.

In Chapter 5, we introduced tearable image warping, an innovative variant of image warping that holds several advantages over pure image warping. Capitalizing on the idea that only part of an object is connected to its physical environment, tearable image warping only maintain semantic connectedness when needed and allows an object in an image to be partially detached from its original background. This approach reduces warping distortion by distributing warping to a wider area of an image and capacitates change in subject-background relationship while preserving scene consistency, making it an effective tool for content-aware image retargeting and image recomposition. For image retargeting, this approach significantly reduced distortion compared to pure warping (Liu et al. 2010), particularly for extreme retargeting cases and is able to preserve semantic connectedness such as shadow and ripples which oftentimes can be violated in results of scene carving (Mansfield et al. 2010). For image recomposition, to our best knowledge, tearable image warping is the first image operator that can produce an effect analogous to change of viewpoint without semantics violation, making it a powerful recomposition tool. With this capability, we can effectively apply geometric transformation to enhance the visual dominance of the photo subject. Combining the subject dominance energy with a set of aesthetics-distance energy

based on selective photographic rules, our recomposition approach successfully enhanced the aesthetics quality of an image. Empirical studies performed on human subjects demonstrate the effectiveness of our approach in enhancing both the subject dominance and aesthetics of a given image.

In summary, we have achieved the threefold objectives of this thesis; (1) to develop **saliency-based aesthetics evaluation** models for aesthetics class and score prediction, (2) to develop a **saliency-based, aesthetics-driven low-level image enhancement** method through saliency retargeting and (3) to develop a **saliency-based, aesthetics-driven image recomposition** method to enhance subject dominance and image aesthetics using the tearable image warping approach.

6.2 Future Research Direction

The research work of this thesis demonstrates that a saliency-based approach that gives core focus to the photo subject(s) in an image, can be an effective strategy for evaluating and enhancing image aesthetics. Here, we identify some possible future research directions that can further enhance the accuracy and effectiveness of our proposed image evaluation and enhancement approaches.

While a subject-focused approach significantly increases the accuracy of aesthetics evaluation, particularly aesthetics classification, using a generic model for all categories of images may be a potential limitation for further accuracy improvement. Different image categories often desire different aesthetics elements. For example, portrait and macro images require low depth of field but in contrast, landscape images often strive for high depth of field. On the other hand, emphasis on the photo subject in portrait and macro images is much greater compared to

landscape images where emphasis is put more on the harmonious combination and composition of many different components in an image. Therefore, a promising future direction is to use a category-based approach that employs category-based features and feature weights in training aesthetics evaluation models, particularly for aesthetics score prediction model. To date, aesthetics score prediction models has yet to achieve high correlation scores. Unlike in aesthetics classification, images with full score range is used for training a score prediction model, making it more sensitive to features used in the model training. We believe that applying a category-based approach on top of our subject-focused approach will generate more precise and relevant aesthetics feature for each image category and could potentially be the key to unlock the bottleneck for better score prediction accuracy.

With the availability of stereo cameras, a promising future direction is to extend our image enhancement approaches to capitalize on the readily available stereo images. Disparity maps obtained from the stereo images can be used to infer the depth order of objects or background and thus ease off user input in both saliency retargeting and tearable warping-based image retargeting and recomposition. In addition to easing off user input, the depth maps also allow the saliency retargeting algorithm to make more gradual and impactful non-uniform sharpness changes to the background of the resulting images. This stereo-enhanced approach is likely capable to produce resulting images with more natural and realistic depth-of-field effect, that potentially leads to better aesthetics experience.

Although tearable image warping produces much less distortion compared to pure warping approaches, warping distortion is largely unavoidable in images with heavy geometric elements in the background, particularly for indoor scene. Thus, another possible future work is to extend tearable image warping to achieve

geometrically consistent image retargeting and recomposition. In addition to preserving geometric consistency, enhancing the aesthetics of indoor scene through desired perspective transformation could be an interesting area of research. Finally, another area of image recomposition that is worth exploring is interactive background warping. The current implementation of interactive background warping is pretty simple and may not be robust enough for images with more complicated background. More innovative and robust user interfaces that enable users to creatively modify the background of an image would be much desirable.

Finally, another potential future work is to extend our tearable image warping algorithm for video retargeting and recomposition. However, the extension to video is non-trivial. As our approach requires image segmentation, one main challenge is to accurately track the object segments across the video frames. In addition, to avoid flickering or waving artifacts, it is crucial to ensure temporal coherence. Extra constraints may be needed to ensure adjacent frames are warped in a coherent manner.

Publication List

Wong, L.K., and Low, K.L. (2009): Saliency-enhanced image aesthetics class prediction. In IEEE International Conference on Image Processing (ICIP), pp. 997-1000.

Wong, L.K., and Low, K.L. (2011): Saliency retargeting: An approach to enhance image aesthetics. In IEEE Workshop on Applications of Computer Vision (WACV), pp. 73-80.

Wong, L.K., and Low, K.L. (2012): Tearable Image Warping for Extreme Image Retargeting. In 30th Computer Graphics International Conference (CGI), pp. 1-8.

Wong, L.K., and Low, K.L. (2012): Enhancing Visual Dominance by Semantics-Preserving Image Recomposition. In 20th ACM International Conference on Multimedia (MM), pp. 845-848.

Bibliography

Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S. (2009), Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1597-1604.

Avidan, S., and Shamir, A. (2007), Seam carving for content-aware image resizing. *ACM Trans. on Graphics*, 26(3), Article 10.

Bae, S., Paris, S., and Durand, F. (2006), Two-scale Tone Management for Photographic Look. *ACM Trans. on Graphics*, 25(3), 637 – 645.

Bae, S., and Durand F. (2007), Defocus Magnification. *Computer Graphics Forum*, 26(3), 571-579.

Banerjee, S. Evans, and B. L. (2007), In-Camera Automation of Photographic Composition Rules. *IEEE Trans. on Image Processing*, 16(7), 1807-1820.

Barnard, K., Cardei, V., and Funt, B. (2002), Comparison of Computational Color Constancy Algorithms-Part 1: Methodology and Experiments with Synthesized Data. *IEEE Trans. on Image Processing*, 11(9), 972-983.

Bhattacharya, S., Sukthankar, R., and Shah, M (2010), A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM International Conference on Multimedia (MM)*, pp. 271-280.

Berkeley Segmentation Dataset Images (BSDS), http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html

Coe, K. (1992), Art: The replicable unit - An inquiry into the possible origin of art as a social behavior. *Journal of Social and Evolutionary Systems*, 15(2), 217-234.

Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., and Xu, Y. (2006), Color harmonization. In *ACM Trans. On Graphics*, 25(3), 624 – 630.

Criminisi, A., Perez, P., and Toyama, K. (2004), Object removal by exemplar-based inpainting. *IEEE Trans. on Image Processing*, 13(9), 1200-121.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006), Studying Aesthetics in Photographic Images Using a Computational Approach. In *European Conference on Computer Vision (ECCV)*, pp. 288-301.

Dong, W., Zhou, N., Paul, J.-C., and Zhang, X. (2009), Optimized image resizing using seam carving and scaling. *ACM Trans. on Graphics*, 28(5), 1-10.

Freeman, M. (2007), The Photographer's Eye: Composition and Design for Better Digital Photos. Lewes:Ilex.

Gasparini, F., Corchs, S., and Schettini, R. (2007), Low-quality image enhancement using visual attention. *Optical Engineering*, 46(4), 1-3.

Gill., P., Murray, W., and Wright, M. (1981) Practical optimization. New York: Academic.

Grant, M., Boyd, S., Ye, and Y., CVX: Matlab software for disciplined convex programming. http://www.stanford.edu/~boyd/cvx.

Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit., S., and Anderson, C.H. (1994), Overcomplete Steerable Pyramid Filters and Rotation Invariance. In

IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 222-228.

Guo, Y., Liu, F., Shi, J., Zhou, Z,-H., and Gleicher, M. (2009), Image retargeting using mesh parametrization. *IEEE Trans. on Multimedia*, 11(5), 856-867.

Harel, J., Koch, C., and Perona, P. (2006), Graph-Based Visual Saliency. In *Conference on Neural Information Processing Systems (NIPS)*, Poster.

Hoiem, D., Efros, A., and Hebert, M. (2007), Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1), 151-172.

Hou, X., and Zhang, L. (2007), Saliency Detection: A Spectral Residual Approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1-8.

Itti, C., Koch, E. and Niebur (1998), A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20 (11), 1254-1259.

Jin, Y., Liu, L., and Wu, Q. (2010), Non-homogeneous Scaling Optimization for Realtime Image Resizing. *The Visual Computer*, 26(6-8), 769-778.

Kao, H.C., Ma, W.C. and Ming, Q. (2008), Esthetics-based quantitative analysis of photo composition. In *Pacific Graphics Conference*, Posters.

Kashyap, R. L. (1994), A robust variable length nonlinear filter for edge enhancement and noise smoothing. *Signal Processing*, 9(3), 505-510.

Koch, C., and Ullman, S. (1985), Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4(4), 219-227.

Levin, A. , Lischinski, D., and Weiss, Y. A Closed Form Solution to Natural Image Matting (2008). In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 228 - 242.

Lind, R. W. (1980), Attention and the Aesthetics Object. *Journal of Aesthetics and Art Criticism*, 39(2), 131-142.

Liu, L., Chen, R., Wolf, L. and Cohen-Or, D. (2010), Optimizing Photo Composition. *Computer Graphics Forum*, 29(2), 469–478.

Liu, L., Jin, Y., and Wu, Q. (2010), Realtime aesthetic image retargeting. In *Eurographics Workshop on Computational Aesthetic in Graphics, Visualization, and Imaging*, pp.1-8.

Ma, L. L., and Zhang, H. J. (2003), A user attention model for video summarization. In *ACM International conference on Multimedia (MM)*, pp. 374-381.

Manos, J.L., and Sakrison, D. J. (1974), The Effects of a Visual Fidelity Criterion on the Encoding of Images. *IEEE Transactions on Information Theory*, 20(4), 525-535.

Mansfield, A., Gehler, P., Van Gool, L., and Rother, C. (2010), Scene carving: Scene consistent image retargeting. In *European Conference on Computer Vision (ECCV)*, pp. 143-156.

Moroney, N. (2000), Local color correction using non-linear masking. In *IS&T/SID Eight Color Imaging Conferenc*e, pp. 108-111.

Nishiyama, M., Okabe, T., Sato, Y., and Sato, I. (2009), Sensation-based photo cropping. In *ACM International Conference on Multimedia (MM)*, pp. 669-672.

Polesel, A., Ramponi, G., and Mathews, V.J. (2000), Image enhancement via adaptive unsharp-masking. *IEEE Trans. on Image Processing*, 9(3), 505-510.

Rahman, Z., Jobson, D., and Woodell, G. (2004), Retinex processing for automatic image enhancement, *Journal of Electronic Imaging*, 13(1), 100-110.

Rizzi, A., Gatta, C., and Marini, D. (2003), A new algorithm for unsupervised global and local color correction, *Pattern Recognition Letters*, 24, 1663-1677.

Rother, C., Kolmogorov, V., and Blake, A. (2004), GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Trans. on Graphics*, 23(3), 309–314.

Rubinstein, M., Shamir, A., and Avidan, S. (2008), Improved seam carving for video retargeting. *ACM Trans. on Graphics*, 27(3), 1-9.

Rubinstein, M., Shamir, A., and Avidan, S. (2009), Multi-operator media retargeting. *ACM Trans. on Graphics*, 28(3), 1-11.

Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., and Cohen, M. (2006), Gaze-Based Interaction for Semi-Automatic Photo Cropping, In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 771-780.

Setlur, V., Lechner, T., Nienhaus, M., and Gooch, B. (2007), Retargeting Images and Video for Preserving Information Saliency. *IEEE Computer Graphics and Applications*, 27(5), pp 80-88.

Scruton, R. (1983), Representation and Photograph, The Aesthetics Understanding, London: Routledge.

Su, S., Durand, F., and Agrawala, M. (2005), De-Emphasis of Distracting Image Regions Using Texture Power Maps. In *ICCV Workshop on Texture Analysis and Synthesis*, pp. 119-124.

Suh, B., Ling, H., Bederson, B., and Jaobs, D. (2003), Automatic thumbnail cropping and it's effectivness. In *ACM Conference on User Interface and Software Technolgy (UIST)*, pp. 95–104.

Teorex Developing Ideas, Inpaint Software. http://www.theinpaint.com/.

Tomasi, C., and Manduchi, R. (1998), Bilateral Filtering for gray and color images, In *IEEE Conference on Computer Vision (ICCV)*, pp. 836-846.

Tong, H., Li, M., Zhang, H., He, J., and Zhang, C. (2002), Classification of Digital Photos Taken by Photographers or Home Users. In *Pacific-Rim Conference on Multimedia*, pp. 367-376.

Vaquero, D., Turk, M., Pulli, K., Tico, M. and Gelfand, N. (2010), A survey of image retargeting techniques. In *SPIE Applications of Digital Image Processing XXXIII*, 779814, doi:10.1117/12.862419.

Wang, J. and Cohen, M. F. (2007), Optimized Color Sampling for Robust Matting. In IEEE Computer Vision and Pattern Recognition (CVPR), pp. 1-8.

Wang, J.Z., Li, J., and Wiederhold, G. (2001), SIMPLIcity: Semantics-Sensitive Intergrated Matching for Picture Libraries. *IEEE Trans. on Pattern Analysis and machine Intelligence*, 23(9), 947-963.

Wang, Y.S., Tai, C.L., Sorkine, O., and Lee, T.Y. (2008), Optimized scale-and-stretch for image resizing. *ACM Trans. of Graphics*, 27 (5), Article 118.

Witten, I. H., and Frank, E. (2005), Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

Wollen, P. (1978), Photography and Aesthetics. *Oxford Journal of Arts*, 4(19), 9-28.

Yan, K., Tang, X., and Jing, Fe. (2006), The Design of High-Level Features for Photo Quality Assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 419-426.

Yang, A.Y., Wright, J., Ma, Y., and Sastry, S.S. (2008), Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110 (2008), 212–225.

Yousef, M., and Hussien, K.F. (2011), ParXII: Optimized, data-parallel exemplar-based image inpainting. In *ACM International Conference on Computer Graphics and Interactive Techniques (ACM SIGGRAPH)*, Poster, ISBN 978-1-4503-0971-4.

Zhang, M., Zhang, L., Sun, Y., Feng, and L. Ma, W. (2005), Auto cropping for digital photographs. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 438-441.