Where2Stand: A Human Position Recommendation System for Souvenir Photography

YINTING WANG, Zhejiang University
MINGLI SONG, Zhejiang University
DACHENG TAO, University of Technology, Sydney
YONG RUI, Microsoft Research Asia
JIAJUN BU, Zhejiang University
AH CHUNG TSOI, Macau University of Science and Technology, Macau
SHAOJIE ZHUO, Qualcomm Canada Inc.
PING TAN, Simon Fraser University

People often take photographs at tourist sites and these pictures usually have two main elements: a person in the foreground and scenery in the background. This type of “souvenir photo” is one of the most common photos clicked by tourists. Although algorithms that aid a user-photographer in taking a well-composed picture of a scene exist [Ni et al. 2013], few studies have addressed the issue of properly positioning human subjects in photographs. In photography, the common guidelines of composing portrait images exist. However, these rules usually do not consider the background scene. Therefore, in this paper, we investigate human-scenery positional relationships and construct a photographic assistance system to optimize the position of human subjects in a given background scene, thereby assisting the user in capturing high-quality souvenir photos. We collect thousands of well-composed portrait photographs to learn human-scenery aesthetic composition rules. In addition, we define a set of negative rules to exclude undesirable compositions. Recommendation results are achieved by combining the first learned positive rule with our proposed negative rules. We implement the proposed system on an Android platform in a smartphone. The system demonstrates its efficacy by producing well-composed souvenir photos.

CCS Concepts:
• Computing methodologies → Computational photography; Image processing;

Additional Key Words and Phrases: Souvenir photography, Photographic composition, Human-scenery positional relationship, Human position recommendation

ACM Reference Format:
DOI: http://dx.doi.org/10.1145/2770879

This work was supported in part by the the Program of International S&T Cooperation (2013DFG12840), by the National Science Foundation of China 61170142, by the National High Technology Research and Development Program of China (2013AA040601), by the Fundamental Research Funds for the Central Universities, by the Australian Research Council Projects FT-130101457, DP-120103730 and LP-140100569, and by the Fondo para o Desenvolvimento das Ciencia das e da Tecnologia, Macau SAR Grant Number 034/2011/A2. M. Song is the corresponding author.

Author's addresses: Y. Wang, M. Song and J. Bu are with Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; D. Tao is with the Center for Quantum Computation & Intelligent Systems, University of Technology, Sydney, NSW 2007, Australia; Y. Rui is with Microsoft Research Asia; A. C. Tsoi is with the Macau University of Science and Technology; S. Zhuo is with Qualcomm Canada Inc.; P. Tan is with School of Computing Science, Simon Fraser University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2015 ACM. 2157-6904/2015/03-ART39 $15.00
DOI: http://dx.doi.org/10.1145/2770879
Fig. 1. Four souvenir photographs. (a) Well-composed: both the woman and landmark are well captured. (b) Poorly composed: the woman blocks the vanishing point of the image, which destroys the perspective in the original scene. (c) Poorly composed: the Statue of Liberty is directly above the woman, generating a dull image. (d) Poorly composed: the main subject is too small to be discernible, which is an undesirable quality in a souvenir photo.

1. INTRODUCTION

The prevalence of digital cameras and smartphones has resulted in wide and growing enthusiasm for personal photography. Among photos captured by an average photographer, one of the most common is what we call the “souvenir photo”. This photo is usually captured during travel and typically consists of a person in the foreground and a landmark in the background. Capturing and sharing souvenir photos is a cross-cultural hobby that has become popular over the last few years. A critical factor that determines the aesthetic quality of a photo is the position of human subjects in a scene. Experienced photographers follow simple rules of composition, such as the rule of thirds [Krages 2005], to improve this aesthetic quality. In addition, human subjects must be positioned within the view frame such that they do not block essential aesthetic features of the scene, e.g., the main landmark or vanishing point. Fig. 1 shows four souvenir photograph samples. Fig. 1(a) displays a well-composed photo in which both the person and the background landmark (the Taj Mahal), are suitably captured in the picture. By contrast, images in Fig. 1 (b)–(d) are poorly composed. For example, the woman in (b) blocks the vanishing point of the image, which destroys the perspective expressiveness of the picture. The subject in (c) stands immediately below the Statue of Liberty, which produces a rather dull picture. The person in (d) appears too small to produce a high-quality souvenir photo.

Although post-processing methods [Liu et al. 2010b] that enhance the aesthetic qualities of captured images exist, we aim to develop a photographic assistance system to help users capture a well-composed souvenir picture in the very beginning. Our idea is inspired by the recent development of what-you-see-is-what-you-get (WYSIWYG) photography, which is demonstrated in the pioneer work in [Baek et al. 2013]. Our study considers the fact that editing and managing a large collection of photographs is a major challenge for an amateur photographer. Moreover, our study echoes the trend of instant photo sharing, in which a user captures a picture and immediately shares it on social network websites without engaging in tedious post-editing. Therefore, capturing well-composed photos immediately and without the need for post-editing is desirable.

In this study, we develop a photographic assistance system that operates on a mobile phone. When a user (i.e., an amateur photographer) takes a photo, our program determines the optimal position for the human subject within the currently framed scene.
It displays a human-shaped icon in the viewfinder that guides the user in positioning the human subject and composing a high-quality souvenir photo.

First, we assemble a dataset of souvenir photographs containing approximately 1000 portraits. These images are evaluated in an aesthetic study for essential qualities of artistic composition—in particular, the relative position of the human subject to the scenery. Second, we extract human subjects and sceneries from the dataset images and then group these images into several categories according to scenery representations. Then, the statistical distribution of human positions with respect to background scenes is learned for each scenery group in order to optimize the positions of human subjects. In this study, this learned distribution is referred to as the “positive rule”. In addition, we introduce a set of “negative rules” based on common rules of composition in photography designed to eliminate improper positioning in photographs. Finally, the ideal position of a human subject in a framed scene is identified by combining these positive and negative rules.

1.1. Related Works
Composition has long been recognized as an essential feature of photography and expert photographers evaluate the quality of a photo based on specific composition rules. Automatic image cropping is a tool commonly employed to optimize photographic composition, and most current image-cropping methods employ a similar framework [Ke et al. 2006, Luo et al. 2011, Luo and Tang 2008]. They typically involve extracting image features such as color and edge and then evaluating aesthetic qualities according to standard composition rules, such as the rule of thirds, shapes and lines, amputation avoidance, visual balance, and diagonal dominance [Krages 2005]. Other cropping methods are based on model learning from a large dataset [Nishiyama et al. 2009, Zhang et al. 2013, Yan et al. 2013]. Nishiyama et al. [2009] trained a support vector machine to label regions of a photograph as high or low quality and then identify the cropping window that achieved the highest quality score by applying the classifier to all cropping candidates. Zhang et al. [2013] examined positional relationships in superpixels and used them to measure the quality of cropping candidates. Yan et al. [2013] conducted a study in which three expert photographers manually cropped 1000 images and then modeled their changes to identify cropping parameters. After they generated an optimal cropping window, they cropped an image to identify dominant visual elements that are included in a high-quality composition. Sometimes, in order to generate satisfactory results, it is necessary not only to crop the image, but to change the relative positions of major areas in the framed scene as well. Liu et al. [2010a] used image retargeting to rearrange image objects and generate optimal results according to composition rules. Bhattacharya et al. [2010] implemented user-guided object segmentation and inpainting to ensure that a final photograph matched user criteria. Most of these approaches are post-processing methods that enhance an existing photograph rather than means of guiding a user to produce a well-composed photo, which we believe is more useful.

Ni et al. [2013] proposed a camera view-selection system that accepts a wide view or a continuous view sequence as input and generates the best-composed photograph. They collected approximately $10^5$ high-ranked photos from the Internet and determined patch spatial correlation distributions as rules to guide composition. This method was effective for scenery capture but did not consider human-scenery positional relationships or changes to photographic composition when a human subject is added to a scene.

Optimizing the position of a human subject in a photo is a new area of research and, based on our findings, only the exemplar-based method presented in [Zhang et al. 2012] has addressed the matter. This method extracts attention and
geometry compositional features from an input scenery frame and then identifies a reference photograph from a portrait database that is the closest compositional match. The human position and pose represented in the referenced photograph are then recommended to the user. However, the attention and geometry features in the collected portraits may have been altered by the subject in the portrait, thereby affecting the accuracy of similarity measurements. Furthermore, the dataset used contained only 232 portraits. This meant that locating a reference photograph that was adequately similar to the input scenery frame was unpredictable. Thus, the method produced less-than-optimal recommendations.

Portrait synthesis is another related research topic in which a portrait is blended into the foreground with popular landmarks shown in the background. Portrait synthesis has been well researched, but most of the existing studies focus on the process of seamlessly blending the foreground and background [Pérez et al. 2003, Jia et al. 2006, Farbman et al. 2009, Tao et al. 2010]. Recently, several data-driven approaches have been proposed that attempt to identify a suitable background image from a large dataset. However, these methods consider only the consistency of color and texture [Chen et al. 2009] or similarity in the illumination of faces [Bitouk et al. 2008]. The pasted position of the foreground subject in terms of photographic composition has not been studied previously. Our proposed method recommends an optimal position for a human subject in an image containing a landmark, and can be integrated with existing portrait synthesis methods to generate improved results.

The remainder of the paper is organized as follows. Section 2 describes the motivation behind the study. Our investigation of human-scenery positional relationships is introduced in Section 3. Our human position recommendation algorithm is presented in Section 4. Section 5 presents experimental results based on the implementation of our system in an Android platform. In addition, we discuss the limitations of our system. Finally, Section 6 presents concluding remarks.

2. PROBLEM STATEMENT AND MOTIVATION

The objective of this study is to develop a human position recommendation system that can be embedded in a mobile platform. Considering the computational limitations of mobile devices, we adopt a series of simple and effective techniques to optimize the position of a human subject in a given picture. A diagrammatic illustration of the pipeline of our recommendation system is presented in Fig. 2. Input for our system is a single scenery frame (Fig. 2 (a)) that contains only the photographer's desired scenery. The view to photograph derives from either users' photography experience or the method described in [Ni et al. 2013]. Our objective is to identify the optimal stand-
ing position for a human subject, dictated by both the relative size of the subject to a given scene and his or her position within that scene. An aesthetic score is calculated for each candidate position based on the recommendation criteria. By discretizing the size parameters, several score maps are generated. Each score map represents scores for all coordinates of a certain size, as illustrated in Fig. 2 (c), and contains an optimal position based on the highest score. Our system then superimposes a human-shaped icon in the scenery frame at the optimal position, as shown in Fig. 2 (d). In this study, a recommendation for the positioning of only a single subject is examined; recommendations for groups of subjects will be examined in a future study.

Fig. 2 (b) shows the positive and negative criteria that are employed in our system. The positive criteria provides the user with positive guidance, that is, to areas in the given framed scene that are suitable for positioning a subject. It can be learned from a dataset of well-composed portraits. We can use an exemplar-based method to produce positive guidance, as in [Zhang et al. 2012]. This method searches for a reference image containing a scene most similar to that in the target scenery frame and then transfers the human position in the reference image to the target scenery frame. However, the exemplar-based method possesses limitations. If the dataset is too small, a reference image containing a scene that is sufficiently similar is not always available and the recommended position of the human subject may not be suitable for the target scene (see Section 5). However, collecting the images of all possible scenes is, for practical purposes, impossible. In addition, the search time increases in direct proportion to an increase in the amount of dataset, meaning the exemplar-based method is unsuitable for mobile devices. Another option is to train a general model for human-scenery correlation in all portrait images, which can then be used to predict the most ideal human position within a given scene. However, real sceneries are too complex and varied to be represented by a unique model. Therefore, in our study, we first group the collected portraits into several categories according to position and structural information of scenery. Human position distribution is then obtained for each group. What we call “positive rule” states that the recommended human position for a given scenery frame must obey the learned distribution of the corresponding group. Our method does not require locating a similar reference image and only needs to simply determine the group to which the target image belongs. The recommendation based on the positive rule is related to the statistical distribution for the entire group, unlike the exemplar-based method, which only recommends the position of a single referenced image. Our method is therefore more error-tolerant.

Using the positive rule alone is insufficient to generate an acceptable result, because the recommendation for all scenes assigned to a particular group will be the same. The positive rule ignores the unique visual elements of a scene such as its salient regions, the vanishing point, and prominent straight lines. These elements determine the location in which a human subject should not be positioned, otherwise compromising compositional quality. Because learning why certain positions are unsuitable is difficult, we define negative rules based on common rules of composition in photography. These positive and negative rules are then combined to evaluate all candidate positions. Finally, the position with the highest score is considered optimal and recommended to the user.

3. HUMAN-SCENERY POSITIONAL RELATIONSHIP ANALYSIS

The main objective in our research is to study human-scenery positional relationships. We build a dataset of well-composed portraits (see Section 5.1) and divide all sceneries into several groups. The distribution of the human positions is then analyzed for each group.
3.1. Human and Scenery Representation

We first need to extract human subjects and sceneries from portraits in the dataset. In this study, we are only concerned with the standing poses of subjects. Recommendations for other poses (e.g., crouching, sitting, lying) will be examined in future research. We ignore the difference in stature of subjects and use a 2D vector and a scalar to denote a human position, where the 2D vector describes location and the scalar reflects size. An automatic face detector [Viola and Jones 2001] is used to identify the dominant face in each image in the database. The face detection results are then examined and any inaccurate detection results are corrected. The 2D coordinates of the face center $x$ (normalized to $[0, 1]$) are used as coordinates of the human location. In addition, the ratio of the face-to-image size is calculated and the scale $\alpha$ is indicated by the square root of the ratio. In this manner, a 3D feature $H(x^T, \alpha)$ can be used to denote the human position.

Automatically extracting scenery from a portrait is extremely challenging. Human subjects may overlap sceneries in portraits; suggesting that current state-of-the-art segmentation methods such as segmentation transfer [Kuettel and Ferrari 2012] and saliency detection [Cheng et al. 2011] are unable to extract scenery completely and effectively. In our study, in order to extract sceneries from the photos in the database, we developed a marking tool, which can be used to mark sceneries manually by identifying some points along the scenery contours. The first seven columns of Fig. 3 show sample dataset images and their scenery marking results. A special type of scenery exists such as a sea, the sky, or a meadow, that is homogeneous. For this type of scenery, we do not mark the image (see the first row of Fig. 3).

Scenery shapes are variable and a simple shape is not sufficient to represent all sceneries. Instead, a 90D feature $S$ is defined to represent accurately the size, position, and structure of certain scenery. The first 85 elements of $S$ are generated by a four-layer pyramid sampling of the binary mask of a scenery. $S(0), S(1 : 4), S(5 : 20),$ and $S(21 : 84)$ correspond to the first four layers, respectively. For each layer, the scenery mask is divided into $2^{l-1} \times 2^{l-1}$ blocks of the same size, where $l$ is the layer index, as shown in the first row of Fig. 4. We then compute the ratio of the pixels marked as scenery in each block. For the first layer, the total ratio of the pixels marked as scenery in the entire image is set to $S(0)$. For the remaining layers, the $2^{l-1} \times 2^{l-1}$ ratios are rearranged into a row vector and set to the corresponding dimensions of $S$. Thus, $S(0)$ reflects the size of the scenery, while $S(1 : 84)$ roughly describes the shape and location. In addition, five other elements $S(85 : 89)$ are introduced to represent the inner structure of the sceneries. To compute $S(85 : 89)$, the image is first divided into two regions by five different ways, as shown in the second row of Fig. 4. We then compute the difference between the ratios of the pixels marked as scenery in the white and black regions and set the difference to the corresponding dimension of $S$. Take $S(85)$ as an example. We cut the image in half lengthways. Let $R_l$ and $R_r$ denote the ratios of the pixels marked as scenery in the left and right parts, respectively. Then, $S(85) = R_l - R_r$. Elements $S(85 : 89)$ reflect that most of the scenery leans to a specific side of the portrait (to the left or right, top or bottom, inside or outside, etc.). These five elements are critical features of the system because different inner structures of sceneries affect the relative positions of subjects. For example, the two buildings in Fig. 5 (a) and (b) are of similar location and size but possess different left-to-right structures. Therefore, the optimal human positions in the two images are different. These two sceneries are very similar if measured by only the first 85 elements, but they are easily distinguished by $S(85)$. The scenery in Fig. 5 (c) has a frame structure, which has its own composition rules, and $S(87)$ can accurately represent the frame structure. For a homogenous scenery, the extracted feature will be a zero vector.
Fig. 3. Fifteen categories of sceneries and their corresponding human-scenery positional relationships. The first seven columns show example sceneries, manually marked with green, for each category. The second column from the right shows the average of scenery masks. The last column illustrates the human position distribution for each category. When computing the average scenery masks and visualizing the human position distributions, we resize all training images to $400 \times 400$ pixels. This is why the images in the last two columns are square.
3.2. Scenery Clustering

With the human subjects and sceneries extracted from all images in our database and represented as $H = \{H_i | 1 \leq i \leq n\}$ and $S = \{S_i | 1 \leq i \leq n\}$, respectively, we next examine human-scenery positional relationships. Because the sizes, positions, and shapes of sceneries vary and because different scenes follow different compositional rules, producing a general model is impossible. Therefore, we classify the training images into $m$ categories according to their scenery features before separately describing the compositional properties of each category.

Let $C = \{C_i | 1 \leq i \leq m\}$ be the centers of $m$ categories of sceneries. Each center $C_i$ is the average of the sceneries for the $i$th category and has the same dimensionality as the scenery feature $S$. To calculate $C_i$, we must minimize the distance sum of each scenery center and all sceneries that belong to the corresponding category,

$$C = \arg\min_{C} \sum_{i=1}^{m} \sum_{S_j \in C_i} d(C_i, S_j),$$

(1)
where $G_i$ is the set of sceneries in the $i$th category and $d(\cdot)$ denotes the distance between a scenery $S$ and its corresponding scenery category center $C$,

$$d(C, S) = \sum_{t=0}^{89} \omega_t \|C(t) - S(t)\|_p.$$  

(2)

Elements $S(85:89)$ contain the structural information of sceneries, which are critical for scenery classification. We therefore assign these five elements a high weight. In our experiments, $\omega_t = 5$ for $85 \leq t \leq 89$ and $\omega_t = 1$ for the other elements. In our study, we first establish all homogeneous sceneries ($S(0:89) = 0$) as a separate category. The other sceneries are then clustered into $m - 1$ categories by asserting an L2-norm to (1) and solving it using the K-means algorithm [Kanungo et al. 2002]. We constrain the minimum size of each category to 20 and determine the optimal category number using the method described in [Kyrgyzov et al. 2007], i.e., $m = 15$. The clustering result is shown in Fig. 3, in which the second column from the right is the average of scenery masks for each category.

3.3. Human Position Statistics

After all collected portraits have been classified into $m$ groups according to their scenery features, we then analyze the human-scenery positional relationship for each scenery category. Two aspects of human position must be considered here: the optimal standing coordinate and optimal size of the human subject. In reality, the size is related to the coordinate. When the capture view is fixed, the size of the human subject is determined by where the user stands. However, modeling the coordinate and size together using a unique model requires recovery of the real 3D depth of the scene, which is both difficult and time-consuming. Therefore, we assume that the position and size are independent and analyze them separately. This assumption may produce artifacts, which are discussed in Section 5.

We model the 2D coordinates $x$ of human subjects for each scenery group $G_i$. For a given group of sceneries, examination of the training images shows that more than one well-composed position always exists in which a human subject can stand. We can use $G_5$ as an example. Because most sceneries in this group are left-right symmetrical, three suitable positions are available: in the left and right third and in the center of the images (see the fifth row of Fig. 3). Based on this observation, we use a Gaussian mixture model (GMM) to model the human coordinates $x$ in our method. The distribution of $x$ is generated by

$$p(x|\Theta_{G_i}) = \sum_{k=1}^{K_i} \omega_k^x \mathcal{N}(x|\mu_{i,k}^x, \Sigma_{i,k}^x).$$  

(3)

$\Theta_{G_i} = \{\omega_k^x, \mu_k^x, \Sigma_k^x | 1 \leq k \leq K_i\}$ denotes the GMM model, where $\omega_k^x$, $\mu_k^x$, and $\Sigma_k^x$ are the weight, mean, and covariance matrix of the $k$th Gaussian component, respectively, and $K_i$ is the total number of Gaussian components. The density is a weighted linear combination of $K_i$ Gaussian densities,

$$\mathcal{N}(x|\mu_{i,k}^x, \Sigma_{i,k}^x) = \frac{1}{(2\pi)^{d/2} |\Sigma_{i,k}^x|^{1/2}} e^{-\frac{1}{2} (x - \mu_{i,k}^x)^T \Sigma_{i,k}^{-1} (x - \mu_{i,k}^x)}.$$  

(4)

We obtain a maximum likelihood parameter set $\Theta_{G_i}^x$ for the GMM by using the expectation-maximization algorithm, as in [Dempster et al. 1977], and the component number $K_i$ is automatically obtained by minimizing the MDL criterion [Bouman et al. 1997] under an experiential hard constraint, $K_i \leq 5$. The last col-
umn of Fig. 3 shows the statistical distribution of quality composition coordinates for human subjects in each scenery category.

We next consider the suitable human size $\alpha$. Deviating from the coordinates, the distribution of human size is simple and concentrates in $0.04$ to $0.07$. If too large, the human subject overlaps the scenery. If the subject is too small, his or her facial details and expression become unclear. A unique Gaussian model $\Theta_{G_i}^\alpha = \{\mu_i^\alpha, \Sigma_i^\alpha\}$ is adopted to model the human size for each scenery category so that

$$p(\alpha|\Theta_{G_i}^\alpha) = \mathcal{N}(\alpha|\mu_i^\alpha, \Sigma_i^\alpha).$$  \hspace{1cm} (5)

$\Theta_{G_i}^\alpha$ and $\Theta_{G_i}^\alpha$ denote the human-scenery positional relationships that we want to learn from the training dataset. In Section 4, we examine how these relationships can be used to position human subjects in souvenir photographs.

4. HUMAN POSITION RECOMMENDATION

We assume that the user has selected the scene and fixed the view for capture. Our recommendation system then informs the user where to stand in order to generate the best-composed photograph. As previously described, our algorithm predicts the optimal human position according to the established positive and negative rules. The obtained human coordinates and scale distribution determine those areas within the visual space that can be occupied by human subjects. This is called the positive rule. By contrast, the negative rules, based on well-known compositional rules, determine improper positions. We combine the positive and negative rules and compute the score for an arbitrary human position $H(x, \alpha)$ by

$$\text{Score}(H) = \text{Score}^+(H) \times \text{Score}^-(H),$$  \hspace{1cm} (6)

where $\text{Score}^+(\cdot)$ and $\text{Score}^-(\cdot)$ denote the scores evaluated by the positive and negative rules, respectively. Only if a position satisfies the positive and negative rules simultaneously will it then be assigned a high score. The position that achieves the highest score is recommended to the user.

4.1. Positive Rule

The positive rule is based on the learned human-scenery positional relationships and determines the location within a scene in which a human subject should stand. Input for the positive rule is a scenery feature $S$. Therefore, we must first separate the desired scenery from the scenery frame. Since the scenery with which the user aims to take photo is always the salient region of the scenery frame, saliency detection-based segmentation methods [Cheng et al. 2011, Achanta et al. 2008] can be applied to segment the scenery. In our study, we applied the region contrast method [Cheng et al. 2011], which computes a saliency value for each superpixel and generates a binary scenery mask using Grabcut. Grabcut is initialized by binarizing the saliency map using a fixed threshold of $0.4$, which ensures that a zero-vector scenery mask is generated for a homogeneous scenery frame.

However, the saliency detection method is not robust when employed with certain visually complex scenes, especially when objects within the scenery contain similar colors and textures as those of the surrounding environment. The saliency detection method cannot determine the user’s desired scenery. Therefore, our system offers two modes of operation: auto and manual. The auto mode segments the scenery automatically by using the aforementioned method, whereas the manual mode requests that the user mark the scenery by hand.

After the scenery mask is obtained, a 90D feature, as defined in Section 3.1, is extracted to represent the segmented scenery, denoted as $S$. By comparing the distances between $S$ and the $m$ cluster centers $C$, the scenery category $G_S$ having the shortest
distance is used as the category for $S$. The score derived from the positive rule can be calculated according to the coordinate and scale distributions of $G_S$,

$$Score^+(H) = p(H|G_S) = p(x|\Theta^{x}_{G_S})p(\alpha|\Theta^{\alpha}_{G_S}),$$  \hspace{1cm} (7)

4.2. Negative Rules

The positive rule recommends an identical human position for scenes belonging to the same group. However, different scenes have unique characteristics that dictate unsuitable regions for human position. Therefore, we propose certain negative rules derived from three standard rules of composition in photography [Zucker 2007, Ziser 2010, Hoffman 2013], as follows:

— The human subject should be positioned such that he or she covers as little as possible of the salient regions of the scenery frame.
— The human subject should not cover the vanishing point of the scene.
— Prominent straight lines should not penetrate the subject’s head.

Each criterion yields a quantitative evaluation for an arbitrary position and the total score of the negative rules is the product of the three separate scores,

$$Score^-(H) = Score^-(H|S) \cdot Score^-(H|V) \cdot Score^-(H|L),$$  \hspace{1cm} (8)

where $S$, $V$, and $L$ denote salient regions, the vanishing point, and straight lines, respectively.

**Salient regions** are most often the scenery itself or relatively important parts of the scenery. Therefore, an evaluation of a human position $H$ must quantify the proportion of the salient regions covered by the human subject,

$$Score^-(H|S) = 1 - \lambda_S \frac{\|Rect_H \& S\|}{\|S\|},$$  \hspace{1cm} (9)

where $S$ is the binary mask of the salient regions, generated by [Cheng et al. 2011], $Rect_H$ denotes the rectangular region occupied by the subject when standing at position $H$, and the symbol “$\&$” represents the AND operation, which computes the salient regions covered by the subject. In addition, $\|\|$ represents the number of nonzero pixels in the binary mask. We utilize a parameter $\lambda_S$ to adjust the range of $Score^-(H|S)$, which serves as the weight of the first negative rule. In our experiments, $\lambda_S$ is set to 4.

The **vanishing point** reflects the gradual change of depth and the perspective expressiveness of the image, which should be retained in the final souvenir photo. To evaluate human position $H$, we need to determine whether the human subject covers the vanishing point,

$$Score^-(H|V) = 1 - \lambda_V \times 1(\text{Rect}_H(V) = 1),$$  \hspace{1cm} (10)

where $V$ denotes the 2D coordinate of the vanishing point, which is detected by the method in [Rother 2002]. The weight $\lambda_V$ is set to 0.9.

**Lines** that penetrate the subject’s head diminish the visual appeal of a portrait. For example, if a flagpole or tree branches appear to project into or through a subject’s head, a strange and unnatural appearance is produced. We apply the Hough transform [Ballard 1981] to detect long straight lines in the scenery frame, and the binary mask is represented by $L$. If a subject’s position is known, we can easily estimate a rough area for the position of the head, denoted as $\text{Head}_H$. The third negative rule requires that interaction of the subject’s head and line mask is empty so that

$$Score^-(H|L) = 1 - \lambda_L \times 1(\text{Head}_H \& L).$$  \hspace{1cm} (11)

In this study, the weight $\lambda_L = 0.7$. 

ACM Transactions on Intelligent Systems and Technology, Vol. 9, No. 4, Article 39, Publication date: March 2015.
Fig. 6. Negative rules and their effects on position recommendation. The results of the negative rules for salient region, vanishing point, and lines, respectively, are shown in the top, middle, and bottom rows. The score maps are generated with all possible \( x \) and a fixed scale \( \alpha = 0.05 \).

Fig. 6 shows the three negative rules and their effects on the recommendation results.

4.3. Optimal Position Searching

The total score for a candidate position is calculated by multiplying the positive and negative scores, as shown in (6). In our experiments, we generated several positive score maps for each scenery group during the learning step. Each score map records positive scores for all coordinates of a fixed size. To compute the positive score for a given scene, we then must simply identify its scenery group and load the positive score maps for that group. For negative scores, we employ image filtering to accelerate the process rather than compute (8) for each human position. To compute the negative score of salient regions, the box filter in [Crow 1984] is applied to the saliency mask, which is equal to calculating \( \| Rect_H \& S \| \) for all \( x \) with a fixed \( \alpha \). The filtering result can be considered as the influence map of salient regions. The influence maps of the vanishing point and straight lines can be generated by using a dilate filter. The negative scores can then be computed as \( 1_{m \times n} - \min(1, \lambda N \times F_N) \), \( N \in \{ S, V, L \} \), where \( 1_{m \times n} \) is a matrix of ones that has the same size as the scenery frame, and \( F_N \) denotes the influence maps. Fig. 7 shows the positive, negative, and total score maps of three sample scenery frames and compares the recommendation results generated with and without the negative rules. We discovered that the results generated by only the positive rule do not consider special visual objects in the scene. This figure demonstrates that the proposed negative rules are necessary even when we have obtained a perfect scenery mask.

We use complete searching to identify the optimal human position. The horizontal coordinate is searched from 0 to 1, and the range of \( \alpha \) is reduced to \([0.04, 0.07]\) based on our experiments. In order to establish a reasonable range for the vertical coordinate, we examined well-composed portraits in the dataset and discovered that nearly all vertical coordinates for human feet are greater than 0.9. For most images, the ground only exists in the bottom 10% of the image. Although the ground may extend from the bottom to the middle of some images, the vertical coordinates for feet should not be less than 0.9. Otherwise, the person will appear extremely small. Based on this observation, the vertical coordinate is searched among \([0.9 - t, 1]\), where \( t \) is the vertical
Fig. 7. Human position recommendation results. (a) Input image, in which the scenery is masked in green. (b) Positive score map. (c) Negative score map. (d) Total score map. (e) Optimal human position computed without negative rules. (f) Optimal human position computed with negative rules. Note that all score maps are generated with a fixed scale $\alpha = 0.5$.

distance from the center of the face to the bottom of the human figure. We traverse the 3D searching space of $H$ with a constant step length of 0.005 to identify the most optimal human position. A human-shaped icon is then superimposed on the scenery frame at the most optimal human position, as shown in Fig. 7. The actual human subject can then enter the scene and stand at the recommended position through the guidance of the icon, and the souvenir photograph can then be captured. If the distance between the recommended coordinates and bottom of the scenery frame is less than the length of the human figure, a half-length icon will be rendered to suggest that the user produce a half-length portrait.

5. EXPERIMENTS AND RESULTS

5.1. Data Collection

We assembled a dataset of well-composed portraits in order to examine human-scenery positional relationships. A large number of portraits were collected by the following ways: (a) pictures were downloaded from common image websites such as Flickr and Picasa using the tags “landscape”, “face”, and “travel”; (b) photographs from social network website albums such as those in Facebook and Renren were downloaded; and (c) the Google image engine was searched using the same keywords as in (a).

Not all images downloaded from the Internet were useful for examination. We asked three undergraduate students to screen the images using the following rules: (a) images should contain only a single dominant person because, in this study, we only analyze the relationship between one person and the scenery; (b) they should contain some type of scenery such as a sea, statue, monument, or an architecture. Images containing only faces or human bodies were removed from the dataset; (c) unnatural or pseudo-images such as cartoons, paintings, or artificially synthesized images were
removed. After the dataset was filtered, the number of photographs was reduced to approximately 3000.

Most of these 3000 images were captured by amateur photographers. Therefore, we conducted an aesthetic study to select the best-composed images by inviting 20 undergraduate students enrolled in an image-processing course and five professional photographers. The undergraduate students were unpaid but received course bonuses. Half of the students were themselves amateur photographers with experience taking portrait photographs. The other half possessed little knowledge of photography. All students majored in digital media art and had completed a sufficient number of art courses that qualified them to judge images for visual aesthetic qualities. The five professional photographers had expert knowledge of photography, which ensured the accuracy of the aesthetic study. Different instructions were given based on the expertise of participants, in photography. Participants with no knowledge of photography were asked to rate the portraits by assuming they were themselves subjects in the photos. Participants who are amateur photographers were asked to consider whether they would recommend that a friend stand in the same position as that in the examined photo. Finally, participants who are professional photographers were asked to provide professional judgment about human subject positioning in the examined photos. Participants rated each image with a score between 1 and 5 (5 representing the highest mark and indicating the best composed photo) and were encouraged to use a full range of scores. All participant evaluation scores for each image were fused by an equally weighted average. The 3000 images were then ranked based on their final scores and the top 500 were selected for inclusion in the portrait dataset. The mean rating for all 3000 images was 2.52 and those images that achieved a rating of 3.68 or higher were defined as “well-composed”. Fig. 8 shows portraits with three levels of ratings. The first row shows well-composed images that were included in the final dataset. Images in the second row attained a score greater than 2.52, meaning they presented the human subject and scenery clearly but violated some composition rules. Images in the third row possessed serious compositional errors and thus earned low scores.

The human-scenery positional relationship should be symmetrical. Therefore, if an image that earned a high score is flipped horizontally, the reversed image should earn the same or similarly high score. Based on this fact, we increased the size of the dataset by adding horizontally flipped versions of original images. The final dataset thus contained 1000 well-composed portraits.
Table I. Mean distances between the recommended coordinates and ground truth over the three datasets.

<table>
<thead>
<tr>
<th></th>
<th>Training Dataset</th>
<th>Extended Dataset #1</th>
<th>Extended Dataset #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>0.1916</td>
<td>0.1810</td>
<td>0.1901</td>
</tr>
<tr>
<td>two</td>
<td>0.1326</td>
<td>0.1352</td>
<td>0.1156</td>
</tr>
<tr>
<td>all</td>
<td>0.0898</td>
<td>0.0908</td>
<td>0.0951</td>
</tr>
<tr>
<td>final</td>
<td></td>
<td></td>
<td>0.0840</td>
</tr>
</tbody>
</table>

Note: The coordinates are normalized to [0,1] for calculating the distances.

5.2. Evaluation and Comparison

To demonstrate the effectiveness of our method, on one hand, we quantitatively evaluated the positive rule based on the learned human-scenery positional relationships on images in the training dataset. On the other hand, we evaluated the proposed recommendation system on various types of scenes, in comparison with the exemplar-based method [Zhang et al. 2012].

5.2.1. Quantitative Evaluation. We first evaluated the positive rule on the training dataset. Each dataset image was assigned to a scenery group $G_i$, and the mean of the $K_i$ GMM components were recommended based on the positive rule. The $K_i$ recommended coordinates were ranked according to their positive scores. We considered the human position in each image in the training dataset as ground truth and calculated the distance between the first recommended coordinate and ground truth. The cumulative histogram of distances is shown as the red curve in Fig. 9 (a). Because many scenes are left-right symmetrical, two optimal human positions are possible. Therefore, recommending only the first coordinate may result in numerous errors. We next calculated the minimum distances between the best two coordinates and ground truth, which decreases considerably, as shown by the green curve. In addition, the minimum distance of all recommended coordinates was calculated and is shown as the blue curve. The mean distances for recommending the first, second, or all coordinates are listed in the second column of Table I.

To further evaluate the positive rule, we apply the same evaluation approach on an extended dataset. The newly added samples in the extended dataset are 200 well-composed souvenir photos collected in the same way as Section 5.1. Note that the extended dataset was only employed for evaluation, and the positive rule was still learned from the original training dataset (1000 images). The evaluation results (Fig. 9 (b) and the third column of Table I) on the extended dataset are similar with those on the training dataset, which demonstrate the effectiveness of our positive rule.

The negative rules are not able to recommend an optimal standing position independently, thus, in order to check the performance of the negative rules, we evaluate their combination with the positive rule. The images in the above two datasets contain human subjects, which may affect the detection of visual elements involved in the negative rules and degrade the recommendation results, hence we built the second extended dataset for evaluating the combination of positive and negative rules. 500 scenery frames were randomly selected from our testing dataset, which don’t contain human subjects. We invited five professional photographers to browse these scenery frames and mark an optimal human position for each frame according to their professional experience. The marked positions were treated as ground truth, and we calculated the distance between the recommendation results of our method and the ground truth. The evaluation results are presented in Fig. 9 (c) and the last column of Table I, which demonstrate that the recommendation results generated by both positive and negative rules are better than the ones generated by only the positive rule.
Fig. 9. Cumulative histograms of distances between the recommended coordinates and ground truth for the three collected datasets: (a) the training dataset, (b) extended dataset #1 and (c) extended dataset #2. The red curve represents the distance between the first recommended coordinate and ground truth position. The green curve represents the minimum distance between the most optimal two coordinates and ground truth. The blue curve represents the minimum distance among all recommended coordinates. The yellow curve represents the distance between the final recommended coordinate computed by both positive and negative rule and ground true.

5.2.2. Comparison with the exemplar-based method. We compared the recommendation results of our method with those of the exemplar-based method [Zhang et al. 2012]. For a fair comparison, we employed the same saliency detection method [Cheng et al. 2011] to segment the scenery for our method and to extract the attention feature for the exemplar-based method. In addition, our extensive professional dataset was applied to the exemplar-based method in order to locate reference images. The exemplar-based method easily locates similar reference images and generates acceptable results for simple scenes. However, this method generates unsatisfactory recommendations for relatively complex scenes for three reasons: (1) the composition representation used in the exemplar-based method includes only attention and geometry features and not the scenery structure or special elements of the query scenery frame; (2) the main subjects in the training images are always salient, which affects the similarity measurement between the query scenery frame and images in the dataset; (3) if the dataset cannot supply a sufficiently similar reference image, the exemplar-based method cannot generate an acceptable result. Fig. 10 provides examples in which our method outperforms the exemplar-based method.

A user study was conducted to quantitatively compare our method with the exemplar-based method. We chose 100 natural scenery images and applied human position recommendation employing both our method using automatic scenery segmentation and the exemplar-based method [Zhang et al. 2012]. The recommendation results obtained from the two methods were shown side-by-side to each participant and in a similar format as that shown in Fig. 10. The participants were asked to identify the most visually appealing image of the two. If the participants believed the two resulting images possessed similar qualities, they were allowed to select both. Twenty people participated in the user study and each participant evaluated 50 sets of images randomly chosen from 100 sets. Finally, the participants selected 1068 ideal recommendation results, among which 68 were in the case of “both selected”. Our method produced 866 selected results, which is 81.1% of the entire collection. The user study further demonstrates the effectiveness of our method. In addition, the subjects are also asked to point out which aspects drive them to make the choice. We give them several options, including “the rule of thirds”, “visual balance”, “human size”, “covering scenery”, “covering vanishing point”, “lines penetrating human head”, “reality” and “others”. If they would like to, the subjects can choose to select one of them for each case. By analyzing the subjects’ selections, we can roughly conclude the factors affecting the quality of the recommendation results. Most of the results generated by
these two methods obey the rule of third and visual balance. Our results outperform the ones of the exemplar-based methods mainly due to the negative rules, especially “covering scenery”. For the cases that the subjects choose the result of the exemplar-based method, they usually took into account the “reality”, which means whether the coordinate and size match or whether the recommended position is incompatible with the given scene (please refer to Section 5.4 for detail discussion).

Fig. 11 displays additional recommendation results generated by our method using automatic scenery segmentation and reveals that our method performs effectively with most common scenes.

5.3. Mobile App
To enable amateur photographers who use mobile devices with integrated cameras to take advantage of our approach, we have developed a human position recommendation system for mobile platforms called Where2Stand. The Where2Stand app embeds our recommendation algorithm in the camera program of mobile devices. The Where2Stand app includes both auto and manual modes. For the auto model, the system takes only a still scenery frame as input and computes the optimal human standing position. The scenery is segmented automatically by a saliency detection-based segmentation method [Cheng et al. 2011]. For the manual mode, the system asks the user to mark the scenery by hand. The user only needs to draw a cycle along the contour of the scenery, which is not very complicated. The Android version of Where2Stand is complete and can be downloaded from Google Play or our project page1, with work on the iOS app in progress.

Fig. 12 shows the following screenshots of Where2Stand on a Samsung Galaxy S4 phone: (a) scenery capturing screen: the user must select the scenery view and capture a scenery frame; (b) scenery marking screen: to eliminate errors from the automatic foreground segmentation method, the user is asked to manually mark the scenery; (c)

---

1http://eagle.zju.edu.cn/~wangyinting/Where2Stand/

ACM Transactions on Intelligent Systems and Technology, Vol. 9, No. 4, Article 39, Publication date: March 2015.
Fig. 11. Recommendation results of our method using automatic scenery segmentation. A human-shaped icon is rendered at the recommended position in each scenery frame. The green masks represent the sceneries extracted by the saliency detection-based segmentation method. For homogeneous sceneries, nothing is masked in green.
portrait capturing screen: the recommendation system identifies the optimal standing position and presents a human-shaped icon to the user. The photographer then must direct the subject to the recommended position and capture the final portrait. The green lines indicate the edges detected in the scenery frame, which help the user align the scenery to the visual frame; (d) portrait browse screen: the captured portraits are saved and the user can browse through them in the photo gallery. For the auto mode, Step (b) is skipped. Once the user has captured a scenery frame in Step (a), our system renders a human-shaped icon at the recommended human position in Step (c).

The Where2Stand app captures a souvenir photograph in a few seconds, including view selection, optimal human position computation, and final portrait capturing. The auto mode of our app takes 5 to 6 seconds for the entire process of recommendation when used on a standard Android smartphone, including the scenery segmentation and the human position recommendation. The manual mode costs less than 10 seconds

Fig. 12. Tour of the Android version of our human position recommendation app, Where2Stand. (a) Scenery capturing screen. (b) Scenery marking screen. (c) Portrait capturing screen. (d) Portrait browse screen.
Fig. 13. Recommendation results with different scenery masks. For each group, the left image is the input image, in which the scenery mask is green, and the right image is the recommendation result of our method.

for the user to mark the scenery and an additional 3 to 4 seconds to identify the optimal standing position for human subject. Actually, the step of scenery marking takes only 2 to 3 seconds for most general scenes, especially when the scenery is small or has a regular shape, e.g., a statue or an architecture.

5.4. Discussions

The positive rule takes a scenery feature as input. If different areas of a scenery frame are marked as scenery, the scenery frame may be assigned to different scenery categories and the system will generate different recommendation results. Fig. 13 compares the recommended human positions for a scene with different areas marked as scenery and illustrates that the step of scenery extraction considerably affects the results. Therefore, when using the auto mode of our system, a failed scenery segmentation result may degrade the performance of the recommendation method. As shown in Fig. 14 (a), the saliency detection-based segmentation method [Cheng et al. 2011] mistakenly extracts the ice on the mountain as the scenery, thereby generating a weak human position in our system. In view of this, we supply both auto and manual mode in our app. If the automatic segmentation method cannot recognize the real scenery, the users can mark it manually.

There are two other limitations of our method, which indicate directions for future research. First, because our method does not recover scene depth, the recommended human coordinates and size may be mismatched. When a subject stands at the recommended coordinates, his or her actual size may be different from that of the recommendation, as Fig. 14 (b) shows. In this situation, the user must move forward or backward in order to balance the coordinates and size, thus enabling him or her to cover a similar region of the portrait as that of the rendered human-shaped icon. Second, the
Fig. 14. Failure cases. (a) Inaccurate scenery extraction causes our system to recommend a weak human position. The green mask represents the scenery segmented by the saliency detection-based segmentation method. (b) The recommended coordinate and size for the human subject do not match the scenery. This figure is a screenshot of our app. (c) The recommended human position is incompatible with the scene.

recommended human position is sometimes incompatible with the given scene, such as in the sea or air, or perhaps on the roof of a building, as exemplified in Fig. 14 (c). To solve these two problems, we must add depth reconstruction and ground detection modules to our system. Existing depth reconstruction and ground detection methods are time-consuming and insufficiently robust. Therefore, they are not included in this system.

6. CONCLUSION

We present a human position recommendation system for producing well-composed souvenir photos. The proposed system adopts two main criteria in the form of positive and negative rules. The positive rule is learned from a dataset of well-composed portraits, and the negative rules are defined based on three common rules of composition in photography. We develop a photographic assistance system for the Android platform, which uses a still scenery frame as input and produces a human-shaped icon at a recommended position on the scenery frame. The photographer can use this recommendation to direct the photographic subject to the proper position in the scene. In the future, this system will be enhanced to address those limitations described in this study, and the iOS version of Where2Stand will be developed.

REFERENCES


Y. Wang et al.


Received November 2014; revised March 2015; accepted April 2015