SAY CHEESE: PERSONAL PHOTOGRAPHY LAYOUT RECOMMENDATION USING 3D AESTHETICS ESTIMATION

Ben Zhang, Ran Ju, Tongwei Ren, Gangshan Wu

State Key Laboratory for Novel Software Technology
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing University, Nanjing 210023, China
{zhangben, juran}@smail.nju.edu.cn, {rentw, gswu}@nju.edu.cn

Abstract. Many people fail to take exquisite pictures in a beautiful scenery for the lack of professional photography knowledge. In this paper, we focus on how to aid people to master daily life photography using a computational layout recommendation method. Given a selected scene, we first generate several synthetic photos with different layouts using 3D estimation. Then we employ a 3D layout aesthetic estimation model to rank the proposed photos. The results with high scores are selected as layout recommendations, which is then translated to a hint for where people shall locate. The key to our success lies on the combination of 3D structures with aesthetic models. The subjective evaluation shows superior preference of our method to previous work. We also give a few application examples to show the power of our method in creating better daily life photographs.

Keywords: Personal photography, Layout recommendation, Aesthetic estimation

1 Introduction

Photographing is one of the favorite things to do for human nowadays. In stark contrast to the matchless enthusiasm, most people don't master professional photography knowledge well. When seeing an amazing view, people always appear struggling in creating satisfactory photographs. In this paper, we propose to solve the problem from the specific perspective of layout recommendation. As shown in Fig. 1, given a scene as background, our method supplies professional advice for people's layout. The final photographs following our guidance get better aesthetic feelings. For example, the left image shows a best visual balance of the scene elements. The middle image shows a good feeling of symmetry. The right image has an obvious direction feeling corresponding to the depth intensity. In fact, photography aesthetics cannot be concluded in several rules. Many other aesthetic rules which have not been established can be learned by our method.

The challenges for making such professional layout suggestions come from two aspects. First, since a photo is a two-dimensional media, it is difficult to

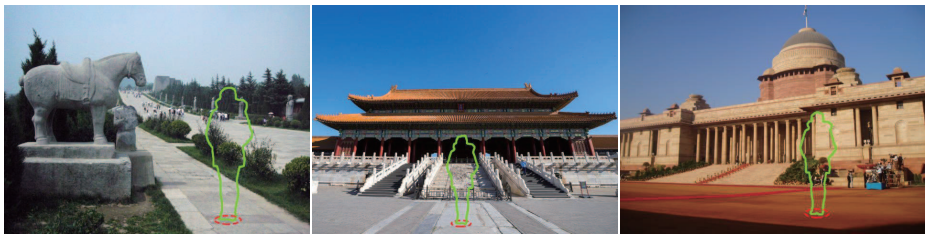

Fig. 1. Illustration of our method. The left image shows a best visual balance of the scene elements. The middle image shows a good feeling of symmetry. The right image has an obvious direction feeling corresponding to the depth intensity.

synthesize reasonable photo proposals with different layouts. Second, previous aesthetic models appear to be less discriminative for these synthetic photos for the lack of consideration of layout, especially 3D structures. To solve the two problems, we propose a novel method works as follow. First, we build human models and detect the ground in order to get reasonable synthetic images. Next, we estimate the 3D layout of the photo using a supervised learning approach based on Markov Random Field (MRF) model. Novel synthetic photos of different layouts are then generated using a grid traversal method. At last, we employ a learning based aesthetic estimation model to rank the synthesized photos with the consideration of photography guideline, color palette, scene layout, 3D structure and elements' saliency. The novel synthetics with top ranking scores are selected and translated to where people shall stand and be visualized on the screen.

To evaluate the effectiveness of the proposed method, we build a novel dataset including 630 photos with rating scores from 12 volunteers. The subjective evaluation results show that the proposed method performs best in the layout recommendation task.

The contributions of this paper are as follows.

- We proposed a novel photo synthesis method based on 3D layout estimation, which is used to generate photography recommendation candidates.
- A novel aesthetic estimation model for photography recommendation is presented, which combines several cues like photography rules, color and scene layout, 3D structure and image elements' saliency.
- We built a dataset designed for the task of making professional photography layout suggestions.

2 RELATED WORK

We briefly review the relevant researches on photo layout recommendation, 3D layout estimation and photo aesthetic estimation as follows.

Photo layout recommendation. Photo layout recommendation is a novel topic proposed in recent years. Many related methods have been proposed but

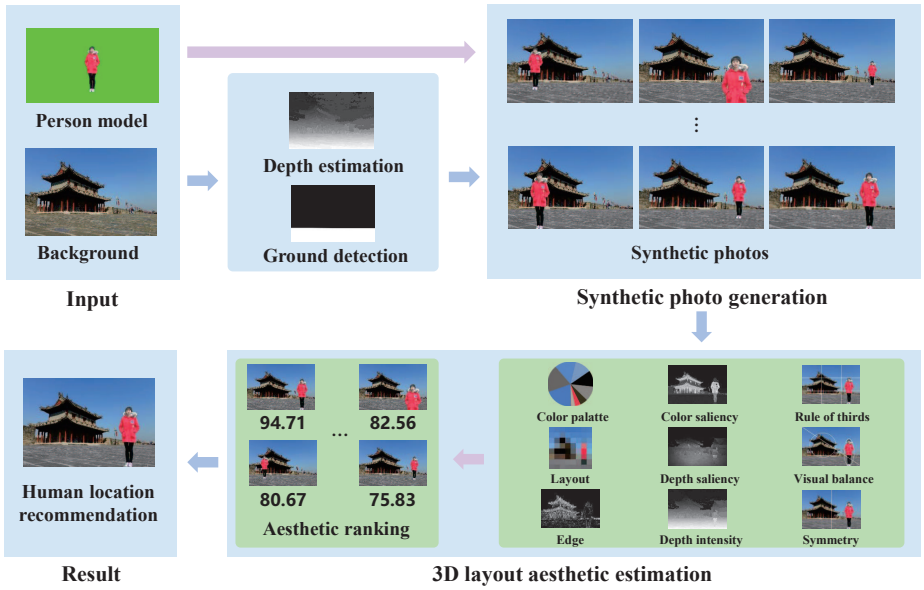

Fig. 2. The framework of our method. Take background and human model as input, ground detection and depth estimation will be executed to generate a set of synthetic photos as shown in the upper right block. Then we combine four features, two saliency intensities and photography guidelines to assess the synthetic photos as shown in the bottom right block. The photo with highest score will be the recommended as shown in the final result.

few of them study the problem of photography layout with human and scene [1][2][3]. Xu *et al.* propose a learning-based method to solve the problem of human position recommendation using 3D point cloud model [4]. However, this work only considers the similarities in social media which limits its application in daily life photography.

3D layout estimation. 3D layout estimation is a challenging problem in computer vision especially for monocular images. Lots of attention has been paid to 3D information extraction and 3D object detection [5][6][7]. Large and specific data are always necessary to support 3D layout estimation from monocular images [8]. In this work we adopt the supervised learning method proposed by Saxena *et al.* to predict the depthmap [9].

Photo aesthetic estimation. Photo aesthetic estimation aims to assess the quality of photos from the aesthetic point of view. Ke *et al.* propose a principled method for designing high level features for photo quality assessment which can be one of the earliest representatives [10]. Later Luo *et al.* propose a content-based photo quality assessment method using regional and global features [11]. Marchesotti *et al.* propose to use generic image descriptors to assess aesthetic quality [12]. Besides, computational rules in photography are utilized for aesthetic assessment in much work [13][14], which are adopted in our work.

3 APPROACH

The framework of our model is shown in Fig. 2. Our application provides a human model package as one of the input elements. For more precise results, the users can set their own model by taking several human photos. When a human model and a background photo are set, the algorithm will call the 3D layout estimation. In this part, we use a discriminatively-trained Markov Random Field (MRF) that incorporates multi-scale for depth inferring [9]. Next, a coarse ground region map is acquired according to the depth intensity and background image [15], which makes the synthetic photos reasonable. After that, we generate several synthetic images with different human figure positions. Finally, we implement a photo aesthetic evaluation combining saliency factors and photography guidelines to rank the candidate images. The position with the highest score will be suggested as a mark region to users.

3.1 3D Layout Estimation

We employ a supervised learning approach by training a set of images with unstructured outdoor environments, which is proposed by Saxena *et al.* to infer the depth for the input image [9]. This model chooses Markov Random Field (MRF) to incorporate local and global features. The algorithm will generate several typical features including haze, texture gradients and variations. The method is trained on images collected from Internet and part of SUN2012 dataset [16], and shows satisfying results to support our work. The depth images are visualized in grayscale where higher intensity indicates nearer and vice versa.

3.2 Human Model and Ground Detection

In our work we assume that only color and posture would slightly influence the final recommendation score. Therefore, we only consider the standing posture and users can choose main color of their own clothes. Meanwhile, users can set their own simple model by taking several human photos for more precise results.

The ground map restricts where people shall stand in synthetic images. We use a Hidden Markov Model (HMM) model with depth intensity to detect coarse ground region [15]. We generate a superpixels map based on the SLIC method [17]. The algorithm will train a HMM model and enact a threshold on depth intensity image. It may not be a necessary part because the wrong layout synthetic images would receive a low ranking in layout aesthetic evaluation. However, this map is a limit condition for synthetic layout generation so that we can improve the efficiency and avoid the obvious ridiculous image like a man standing in the air or on a tree.

3.3 Synthetic Photo Generation

The key of our application is to generate the best synthetic layout images. In our application, we also give an alternative selection on whether salient regions shall be remained. In our method, we consider the human figure to be the candidate

component and the ground as a support plane. The image is separated into $n*n$ grids. We put the geometry center of human figure image at each grid center. The algorithm will traverse the grid image to generate the candidate synthetic images.

In each grid, we compute the depth intensity at the human figure bottom point. Human model will be proportionally resized according to the grid depth:

$$Z = \frac{k}{D} \tag{1}$$

where Z is the coefficient of human model size and D is the depth intensity of current human figure bottom point. k is a parameter to control the basic size of human figure according to the camera parameters. Besides, the human figure will slightly be modified in contrast and luminance corresponding to the background. Along with the $n*n$ grid image, if the depth process has been done and the candidate synthetic images have been generated, we plan to increase the accurate aesthetic of the human figure position. To accomplish that, we shrink the basic grid into $m*m$, and repeat the traversal steps. However, this process is alternative, one condition is that the human figure position is not at the saliency region if the saliency region requested well preserved. The other one is that running time of this model should be limited in a threshold which is set to 0.5 second in our application.

3.4 Assessment for Synthesis Images

Computational Photography Model We adopt many photography guidelines (PG) in our work such as symmetry and patterns, rule of thirds and visual balance [13][18][19]. Rule of thirds (RT) may be the most popular photography guidelines which is based on the golden ratio. The image is split into 9 parts equally with two average horizontal and vertical lines. The main object should be put in the intersection, which is formalized as follows

$$Score(RT) = \max_{j=1,2,3,4} \frac{1}{\sum_{i\in S} d(S_i, I_j)} \sum_{i\in S} d(S_i, I_j) \exp(-\frac{d^2(S_c, I_j)}{2\sigma}) \tag{2}$$

where S is the salient region set, I is the interaction set and c is the center point in salient region. σ is a parameter to control the range of score and set to 0.18.

Visual balance (VB) suggests the visual mass to be put in the center of a scene. We compute the saliency and depth intensity to get the visual center score as follows

$$Score(VB) = \frac{S_i}{\phi D_i} \exp(-\frac{\sum_{i\in S} d^2(S_i, C)}{2\upsilon}) \tag{3}$$

where D is the depth intensity and C is the center point of the image. ϕ is a parameter to control the fusion rate of saliency intensity and depth intensity and set to 0.5. υ is a parameter to control the range of score and set to 0.2.

Symmetry and patterns (SP) is a special rule in our model. If the image get a high score of symmetry, we will intend to put the human figure in the center

which is corresponding to the visual balance. We compute the correspondence split by middle vertical line.

$$Score(SP) = \lambda \exp(-\frac{\sum_{\substack{i+p=H \\ j+q=W}} |(x_i, y_j) - (x_p, y_q)|^2}{2\omega}) \tag{4}$$

where H is the height of the image, W is the width of the image and (x, y) is each point divided by the middle vertical line. λ and ω is parameters to control the range of score and set to 0.7 and 0.18. Finally, we combine those three computational photography guidelines to be a complete model and compute the final score.

Photo Aesthetic Evaluation We use an instance-based approach for photo aesthetic evaluation algorithm [14]. Four features are selected as the main components including layout composition, edge composition, color palette and global texture features. Besides, several features such as blur feature, dark channel, contrasts and HSV counts are taken into consideration.

Furthermore, we employ the 3D structure for our aesthetic evaluation. We arrange the high-quality images in dataset and generate a few high-quality templates in depth intensity form, as well as low quality ones. The difference between high-quality (hq) and low-quality (lq) templates is considered as a unique feature. Totally it is 24 features and a binary classifier by using support vector machine (SVM) is utilized to train our dataset, which is formalized as follows

$$Score(AE) = \frac{w^T x - b}{|w^T x_{hq} - b|} \exp(-\frac{|w^T x - b|^2}{2\mu}) \tag{5}$$

where w^T and b is the coefficient vector trained by SVM, and x is the features defined above. μ is a parameter to control the range of score and set to 0.2.

Saliency Evaluation Except for the aesthetic evaluation mentioned above, we consider a further step to enhance the difference between 'good' and 'bad' synthetic images. As a common sense, people expect figures or objects to be salient in the scene. Accordingly, we consider both color (c) saliency and depth (d) saliency in our aesthetic evaluation model.

We use a saliency method that based on anisotropic center-surround difference [20]. We can get a initial saliency map according to the measure map which is resulted to [0,255]. Besides, we use a soft image abstraction representation method to generate a saliency map based on color intensity [21]. Based on the formal steps, the saliency map is generated with a global uniqueness indicator and a color spatial distribution indicator. We combine this map and the depth based map as complex saliency (CS) component in assessment module as follows

$$Score(CS) = \frac{a_{dm} + a_{cm}}{\xi A} exp(-\frac{\sum_{a \in S}(a_d + a_c)}{2\nu A}) \tag{6}$$

where a is the pixel area of salient objects summed by depth based area a_d and color based area a_c. m is the pixel area of human figure region. A is the total

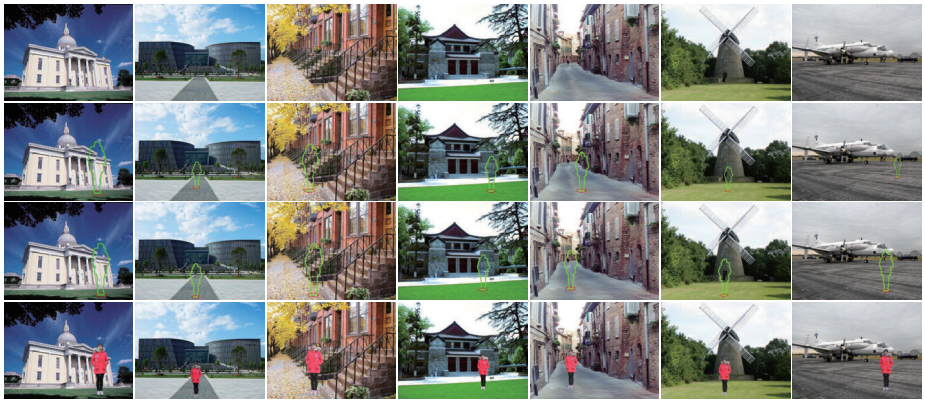

Fig. 3. Several results of our application. The first row is the photography scene. The second row is the professional photographer recommendation. The third row is best results of our method. The fourth row is the abridged view of synthetic images to show an intuitive feeling.

pixel area of the photo. ξ and ν is parameters to control the range of score and set to 0.7 and 0.18.

Finally, we combine three components as one image quality assessment model. We find the dominate images which gets the highest score as follows

$$Dom\{i\} = \underset{i \in synthesis}{\arg\max} \ \alpha CS(i) + \beta AE(i) + \gamma PG(i) \tag{7}$$

where i is each synthetic image. α controls the influence of complex saliency, β controls the influence of aesthetic evaluation model, γ controls the influence of photography guidelines. We linearly set complex saliency, aesthetic evaluation and computational photography guidelines factors with the equal weight 0.33.

4 EXPERIMENTS AND ANALYSIS

4.1 Experimental Settings

To the best knowledge of ours, professional personal photography layout suggestion is a relatively minor area. Thus, we establish a small-scale dataset for our application experiment. We choose 21 scenes as our initial data structure. All of them is outdoor environments with or without salient objects. For each scene, we take a background and three people in the photo with different positions and photography layouts. Moreover, one person is dressed in clothes with different colors to check the color palette effect. One person stands in different positions especially form an obvious unsightly layout. Therefore, we totally collect 630 images as our dataset. And we simply labeled non-person images as 'background', deliberately ugly setting images as 'unsight',and the remains as 'normal'. Moreover, we attach a rating score to each photos which is not used in this paper's

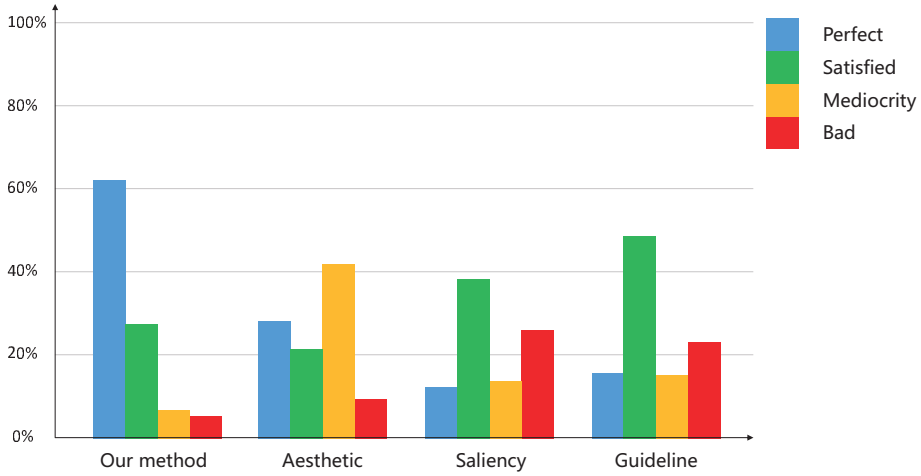

Fig. 4. The comparison of four assessment on user markings: Photography guidelines only method, saliency only method, aesthetic only method and our proposed method

work. This dataset can be used as a basal one for the task of making professional photography layout recommendation.

We invited 4 professional photographers to give their advice on where is the best position to stand. Because of the workload limit, we select 100 photos from our dataset, the Internet and SUN2012 dataset [16] for them and use the professional suggestion as a contrast. Besides, we downsample the photo to 640×480 resolution for unity. Moreover, we arrange a user study to check the accuracy of our method. we design a computational guidelines only assessment, an aesthetic only assessment, a saliency only assessment to show the enhancement and shortcoming in our method. We invited 12 volunteers with various backgrounds including 4 females and 8 males to test those synthetic photos based on saliency only method, aesthetic only method, computational guidelines only method and our combined method. Each photo is requested to mark 'Perfect', 'Satisfied', 'Mediocrity' and 'Bad'. 'Perfect' is the exactly best position in the users' mind. 'Satisfied' is that the current position is not the best but remain good aesthetic. 'Mediocrity' is the photo looks common and user can not say it is beautiful or ugly. 'Bad' is the photo looks ugly and not willing to show them to others. The professional photography contrast will show the disparity between our method and professional photography aesthetic. The user study will show the public view on our method.

4.2 Experimental Results and Discussion

Our experiments are based on the user study and professional photography contrast. From Fig. 4, we can see the great superiority of our method. The accurate data of our 'Perfect' is 62.42% and 'Satisfied' is 26.33%. Therefore, in 88.75%

situation, users think our method is pleasing and applicable. While only 4.92% is bad and when it comes to separate method, it increase a lot. 27.58% saliency only method and 23.08% guideline only method is bad. Nearly half of the aesthetic method is 'Mediocrity' and 'Bad'. Therefore, our method is not just based on one main method. It can not be applicable if only one of those method considered. Moreover, the most satisfied situation is the guideline only method. It shows that guideline is the most common method to make people feel aesthetic about a photo, which is corresponding to the traditional photography rules. The saliency only method shows the worst 'Perfect' rate. But when it is combined by some aesthetic methods, it shows an obvious improvement.

The distance between our method result and professional photographer result shows that most of the distance is less than 60 pixels. Totally it is 400 photos and each block is 20 pixel distance plus from darkness to the brightness. 20 pixels less distance occupies 23.75% and [20,40] pixel distance is 32.5%. Therefore, along with 18.5% of the [40,60] pixel distance, totally 74.75% situation is less than 60 pixel distance. Considering the photo is 640×480 resolution, we can tell that three quarters of our results are similar with professional view. However, about 20% situation is far away from the professional recommendation. Professional photography aesthetic is a complex problem in various scenes, but our method can be applied with most situations and satisfied with amateur photographers.

5 CONCLUSIONS

In this paper, we propose a method to the specific perspective of layout recommendation. Combining 3D structures into aesthetic estimation models along with saliency intensity shows a better discrimination. We also built a dataset designed for the task of making professional photography layout suggestion. According to the scene, our method supplies professional advice for human locations.

Acknowledgments

This work is supported by National Science Foundation of China (61321491, 61202320), Research Project of Excellent State Key Laboratory (61223003), Research Fund of the State Key Laboratory for Novel Software Technology at Nanjing University (ZZKT2016B09), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Bourke, S., McCarthy, K., Smyth, B.: The social camera: a case-study in contextual image recommendation. In: Proceedings of the 16th international conference on Intelligent user interfaces, ACM (2011) 13–22
2. Tian, Y., Wang, W., Gong, X., Que, X., Ma, J.: An enhanced personal photo recommendation system by fusing contextual and textual features on mobile device. Consumer Electronics, IEEE Transactions on **59**(1) (2013) 220–228

3. Elahi, N., Karlsen, R., Holsbø, E.J.: Personalized photo recommendation by leveraging user modeling on social network. In: Proceedings of International Conference on Information Integration and Web-based Applications & Services, ACM (2013) 68
4. Xu, P., Yao, H., Ji, R., Liu, X.M., Sun, X.: Where should i stand? learning based human position recommendation for mobile photographing. Multimedia Tools and Applications **69**(1) (2014) 3–29
5. Xiang, Y., Savarese, S.: Estimating the aspect layout of object categories. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3410–3417
6. Geiger, A., Wojek, C., Urtasun, R.: Joint 3d estimation of objects and scene layout. In: Advances in Neural Information Processing Systems. (2011) 1467–1475
7. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 623–630
8. Kumar, S., Hebert, M.: Discriminative fields for modeling spatial dependencies in natural images, NIPS (2003)
9. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems. (2005) 1161–1168
10. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 1., IEEE (2006) 419–426
11. Luo, W., Wang, X., Tang, X.: Content-based photo quality assessment. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 2206–2213
12. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 1784–1791
13. Liu, L., Chen, R., Wolf, L., Cohen-Or, D.: Optimizing photo composition. In: Computer Graphics Forum. Volume 29., Wiley Online Library (2010) 469–478
14. Lo, K.Y., Liu, K.H., Chen, C.S.: Assessment of photo aesthetics with efficiency. In: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE (2012) 2186–2189
15. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: Computer Vision–ECCV 2014. Springer (2014) 345–360
16. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE (2010) 3485–3492
17. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(11) (2012) 2274–2282
18. Grill, T., Scanlon, M.: Photographic composition. Amphoto Books (1990)
19. Krages, B.: Photography: the art of composition. Skyhorse Publishing, Inc. (2012)
20. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: Image Processing (ICIP), 2014 IEEE International Conference on, IEEE (2014) 1115–1119
21. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 1529–1536