

Aesthetic Guideline Driven Photography by Robots

Raghudeep Gadde and Kamalakar Karlapalem

Center for Data Engineering

International Institute of Information Technology - Hyderabad, India

raghudeep.gadde@research.iiit.ac.in, kamal@iiit.ac.in

Abstract

Robots depend on captured images for perceiving the environment. A robot can replace a human in capturing quality photographs for publishing. In this paper, we employ an iterative photo capture by robots (by repositioning itself) to capture good quality photographs. Our image quality assessment approach is based on few high level features of the image combined with some of the aesthetic guidelines of professional photography. Our system can also be used in web image search applications to rank images. We test our quality assessment approach on a large and diversified dataset and our system is able to achieve a classification accuracy of 79%. We assess the aesthetic error in the captured image and estimate the change required in orientation of the robot to retake an aesthetically better photograph. Our experiments are conducted on NAO robot with no stereo vision. The results demonstrate that our system can be used to capture professional photographs which are in accord with the human professional photography.

1 Introduction

The goal of this work is to get robots to take good photographs that are coherent with humans perception. In this research, we categorize the initially captured photographs into two classes, namely good and bad quality images by assessing their *visual appeal*. We then *recapture (if required) a better photograph*, according to the aesthetic composition guidelines of professional photography by changing the orientation of the robot camera or the part containing camera. A computationally efficient image quality assessment technique and a methodology to estimate the desired change in the orientation is required to recapture an aesthetically better image. The current state of art of image quality assessment needs high processing power [Luo and Tang, 2008]. In this paper, we develop a computationally efficient quality assessment model. We then propose an iterative approach for capturing better photographs.

Our quality assessment work differentiates the high and low visually appealing photographs shown in Figure 1. It is independent of type of subject in the image (for example it

can be an object or a human or a scenery). In this work, we do not deal with parameters associated with the camera like shutter speed, exposure etc., as their values depend on the type of the photograph required. Further we limit ourselves to robots which do not have stereo camera. Our work is also confined to static scenes. It is assumed that the robot (like NAO [Gouaillier *et al.*, 2008]) can rotate the camera or the part containing the camera in all four directions, up, down, left and right.

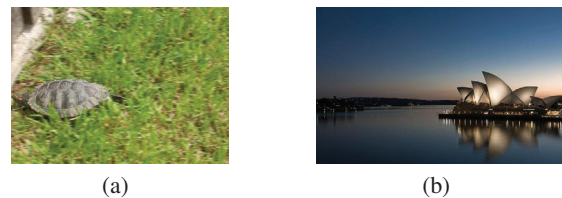


Figure 1: Example images of 1(a) low quality and 1(b) high quality photograph

1.1 Motivation

There are two main advantages of having good photographs taken by a robot, (i) commercially they can be used in robot journalism and for publishing because of the increasing demand for professional photographers, and (ii) having good photographs can help efficiently process the image for decision making by the robot, for example in robot soccer. In addition, robot photography can also be used to take photographs in locations where humans find it hard like in difficult terrains or unreachable places.

Figure 1 shows two photographs. Humans can judge that the left photograph is of low quality and that the right photograph is of high quality, but a robot needs to decipher it. Helping a robot to judge the visual appeal of the captured image is challenging because it is based on combination of features of the image and the aesthetic guidelines of professional photography. Figure 2 shows an example of aesthetically appealing photos. Professional photographers rate the left image of higher quality than the photograph on the right. Our methodology used by the robot to classify images can also be used for other applications like web image ranking.



Figure 2: Example images, 2(a), following the rule of thirds composition guideline of photography, ($f_{th} = 0.11$, $f_{gr} = 2.08$) and 2(b), ignoring it, ($f_{th} = 0.19$, $f_{gr} = 0.87$)

1.2 Related Work

Photographer robot systems like [Byers *et al.*, 2003], [Ahn *et al.*, 2006], [Kim *et al.*, 2010] are predominantly limited to capturing photographs of humans with certain designated compositions based on the approach and the results presented in their papers. They use image processing algorithms like face detection and skin color detection techniques to detect the presence of humans in the scene and capture them. Our approach is generic and does not rely on the subject of the image being captured.

Recent developments in image processing have given rise to several techniques like [Wang *et al.*, 2002], [Tong *et al.*, 2004] for no-reference image quality assessment. The most recent work by [Ke *et al.*, 2006], [Luo and Tang, 2008] extract a set of features on a captured image and compare them with the features of the training data-set containing good and bad images. The features are based on properties of a good professional photograph. According to [Luo and Tang, 2008], they consider an image to be of high quality if its subject has the maximum attention and the absence of regions which distract attention from the subject. They assess the quality of an image by extracting the subject region from the image. The extracted features measure the composition, lighting, focus control and color of the image of the subject region compared to the whole image. Their approach uses the detection of blurred regions in the image to extract the subject region by subtracting the background (blurred regions) from the original image. Their model requires smoothing and convolving with kernels of size $k \times k$, $\{k = 5, 10 \text{ or } 20\}$, approximately 50 times to get better results. Although [Luo and Tang, 2008] claim up to 93% accuracy rate, their approach is computationally intensive. Devices like digital cameras and mobile robots which have less computation power, cannot use these approaches.

Recently computation efficient algorithms like spectral residual (SR) [Hou and Zhang, 2007], phase spectrum of fourier transform (PFT) [Ma and Zhang, 2008] and [Achanta *et al.*, 2009] have been developed to extract the salient regions of an image which in general matches with the subject region of the image by processing it in frequency domain. According to the saliency model comparison study done by Achanta [2009], the SR model is slightly computationally efficient than other models but the model proposed by Achanta, gives better results than SR. In our approach in section 3, we use the saliency model proposed by [Achanta *et al.*, 2009] aided by the features of [Luo and Tang, 2008].

1.3 Contribution and Organization of the Paper

In this paper, we make two major contributions, (i) we present a computationally efficient mechanism to judge the photograph captured by the robot and (ii) a methodology to reorient robots by themselves (if required), to capture better photographs. The remainder of this paper is organized in the following manner. Section 2 describes the properties followed in general of a good photograph. In section 3, we present our image quality assessment approach and a methodology which the robot can employ to reorient itself if required to capture better images like in Figures 2, 4. We evaluate the proposed approach in section 4 and we conclude in section 5.

2 Elements of a Good Photographic Image

A photograph can be assessed based on its major components some of which are light, color, tone and contrast; texture; focus and depth of the field; viewpoint, space and perspective (shape); line, balance and composition [Harris, 2010]. Because of limited features available on the camera of the robot, we use only the light, color and contrast features of an image as proposed by [Luo and Tang, 2008]. Other aspects which are important for a good photograph are visual balance and perspective. Efficient computational models do not exist to find visual balance of an image. Perspective requires that the spatial orientation of the subject of most of the images in general follow the spatial compositional guidelines [Grill and Scanlon, 1990; Lamb and Stevens, 2010] which help to produce balanced images and holds the aimed subject in focus. Figures 2, 4 show some examples. Professional photographers rate Figures 2(a), 4(a) as more visually appealing than their corresponding Figures 2(b), 4(b). A good photographer can follow any of the composition guidelines of professional photography [Harris, 2010; Lamb and Stevens, 2010]. We apply the two well known composition guidelines namely, the rule of thirds and the golden ratio rule. Professional photographs in general have the subject region in focus and the remaining background blurred [Luo and Tang, 2008].

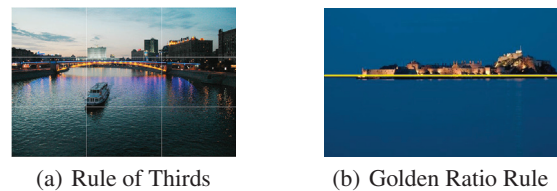


Figure 3: Example images showing the composition guidelines of photography

The Rule of Thirds: According to this rule [Harris, 2010], an image should be imagined as divided into nine equal parts by two equally-spaced horizontal lines and two equally-spaced vertical lines, and that important compositional elements should be placed along these lines or their intersections (*i.e. intersection points*). Aligning a subject with these points creates more tension, energy and interest in the composition than simply centering only the subject. Figure 2 shows an example.

The Golden Ratio Rule: This rule requires the ratio between the areas of the rectangles formed because of the horizon line [Ang, 2004] be equal to the *golden mean*, 1.618, to be more pleasing to the eye. An example is shown in Figure 4.

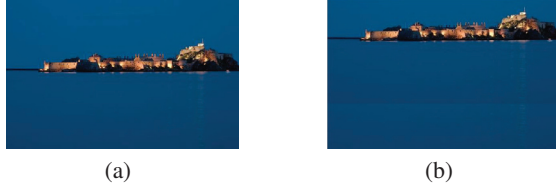


Figure 4: Image examples, Figure 4(a) following the golden ratio composition rule, ($f_{th} = 0.12, f_{gr} = 0.29$) and 4(b) ignoring it, ($f_{th} = 0.17, f_{gr} = 1.61$)

3 Iterative Approach To Robot Photography

In this section, we present our quality assessment approach and the methodology to estimate the change required in its orientation to capture better images for a photographer robot. Figure 5 shows the flow of our approach.

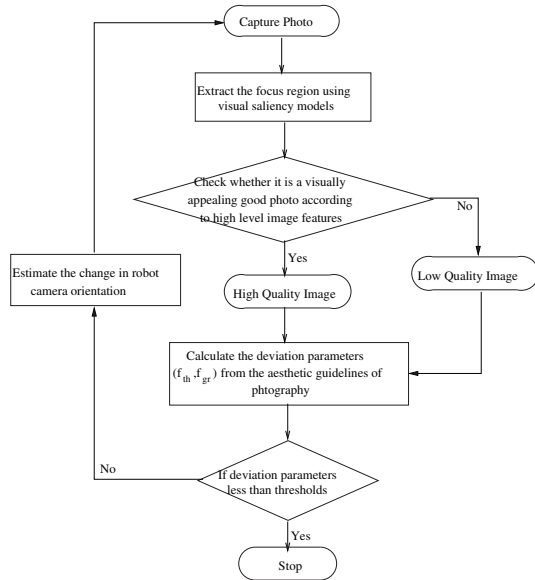


Figure 5: Robot Photography Methodology

The robot captures an image when it is asked to. The visual quality of the captured image is assessed and the desired change in the orientation of the robot camera is determined using the aesthetic deviation readings. A new image is re-captured if the aesthetic parameter readings are larger than certain thresholds. This feedback procedure is repeated until an image with less aesthetic deviation is captured.

3.1 Saliency Based Image Quality Assessment

Our approach to classify the images into high and low quality according to their visual appeal is based on extracting the

focused region directly contrary to the extraction of blurred regions and subtracting them from the original image as followed in [Luo and Tang, 2008]. We use the visual attention model by [Achanta *et al.*, 2009] to extract the salient regions of the image. The generated saliency map is thresholded to extract the focused subject region. The spatial domain features proposed by [Luo and Tang, 2008] and the two aesthetic guidelines of professional photography, the rule of thirds and the golden ratio rule are used to assess the quality of the captured image.

For our experiments the parameter for thresholding the saliency map are decided after a series of experiments on a dataset consisting of good professional photographs. The saliency maps generated are normalized and experiments were performed by varying the threshold. The accuracy rate varied between 75% to 80% for thresholds between 0.5 to 0.75. Figure 6 shows an example of the extracted subject region after thresholding. The extracted region is used to compute the high level features of an image as proposed by [Luo and Tang, 2008] which constitute of the quantitative metrics on subject clarity, lighting, composition and color. These features, namely clarity contrast feature (f_c), lighting feature (f_l), simplicity feature (f_s), color harmony feature (f_h) were developed statistically. These parameters are learned using the basic two class SVM classifier (in Matlab) and run on the captured image to judge its visual appeal (i.e. good or bad quality photograph).

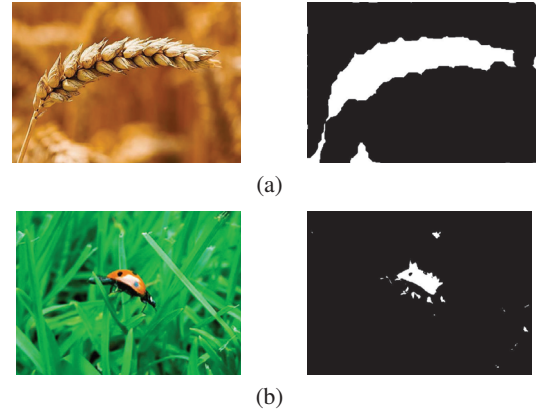


Figure 6: Extracted salient regions on 6(a) high quality image and 6(b) low quality image

The aesthetic guidelines of professional photography are applied on the images which are judged as good images. The rule of thirds feature (f_{th}), is calculated as the minimum deviation observed by the centroid (C_x, C_y) of the extracted subject region.

$$f_{th} = \min_{i=1,2,3,4} \{ \sqrt{(C_x - P_{ix})^2/X^2 + (C_y - P_{iy})^2/Y^2} \}$$

where (P_{ix}, P_{iy}), $i = 1, 2, 3, 4$ are the four intersection points of the image. X and Y are width and height of the image respectively. f_{th} has a bounded range $[0, 0.47]$ as the maximum deviation from rules of thirds occur when the centroid of subject region coincides with any of the corners of the image.

The golden ratio feature (f_{gr}), is calculated by computing the ratio (r) of areas of the rectangles formed by the horizon line of the image which is generated using the vanishing point detector [Leykin, 2006].

$$f_{gr} = |\max\{r, 1/r\} - 1.618|$$

Large values of f_{th} or f_{gr} indicate high aesthetic deviation of the photograph. Note that the composition guidelines are followed if the f_{th} , f_{gr} feature values are less than certain thresholds. These thresholds are determined by taking the average of the corresponding feature values computed on a dataset of good professional photographs. In our experiments, the thresholds set for f_{th} and f_{gr} are 0.25 and 0.30 respectively.

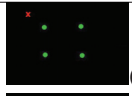
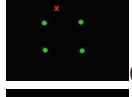
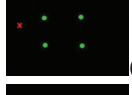
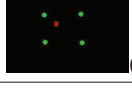
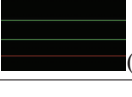
3.2 Robot Re-Orientation

In this section we present an approach to calibrate the change ($\Delta\theta$) required to reorient the robot camera. To satisfy the rule of thirds, the centroid (C_{Rx}, C_{Ry}) of the subject region should coincide with any of the four intersecting points (P_i) as shown in section 2. The point nearest to the centroid region is chosen by calculating the Euclidean distance.

$$distance = \min_{i=1,2,3,4} \{\sqrt{(C_x - P_{ix})^2 + (C_y - P_{iy})^2}\}$$

To shift the centroid of the subject region to the desired location, the orientation of the robot needs to be changed with a certain angle ($\Delta\theta = (\Delta\theta_x, \Delta\theta_y)$) along the axes of the photograph. For example in Figure 2, the camera should be rotated to its left and upwards to make the subject region coincide with the nearest of the four points of the thirds rule.

Table 1: Some possible cases

Cases	(f_{th} , Direction)	Cases	(f_{gr} , Direction)
	(0.242, $\leftarrow\uparrow$)		(2.812, $\uparrow$)
	(0.373, $\rightarrow\uparrow$)		(0.356, $\downarrow$)
	(0.182, $\leftarrow\downarrow$)		(0.343, $\uparrow$)
	(0.107, $\rightarrow\downarrow$)		(2.812, $\uparrow$)

For images following the golden ratio rule, the deviation of the horizon line is calculated using the Manhattan distance of corresponding points on the deviated and the aesthetical horizon lines. In the example, shown in Figure 4 the robot camera should be reoriented in the upwards direction. Table 1 shows the directions in which the robot camera must reorient itself for some possible cases in upper left quadrant. The green

regions are the aesthetically desired locations in the image, while the red are the deviated regions.

A naive approach to reorient the robot could be by changing the orientation of the robot camera in integral multiples of small angle ($\delta\theta$), say 1° . The problem with this approach is the error in the movement of the robot camera gets compounded, which may sometimes result in much more deviated photographs. Also the number of intermediate images captured increases linearly with the deviation.

To reduce the compounded error and reorient the robot in reduced time we follow an approach which is logarithmically converging to capture the required photograph. The following algorithm where (C_x, C_y), (P_x, P_y) are the centroid of subject region and the nearest point from rule of thirds and r is the ratio of areas of upper rectangle to the lower rectangle formed by the horizon line drives the robot re-orientation.

Algorithm 1 Robot Reorientation

```

1:  $\Theta = (Angle\_view\_range\_of\_robot\_camera)/2$ 
2: Direction to reorient is known from table 1
3: while  $f_{th} > 0.25$  or  $f_{gr} > 0.30$  do
4:   if  $C_x - P_x < 0$  or  $r < 0.618$  or ( $r > 1$  and  $r < 1.618$ ) then
5:     Camera should be reoriented  $\uparrow$ 
6:   else
7:     Camera should be reoriented  $\downarrow$ 
8:   end if
9:   if  $C_y - P_y < 0$  then
10:    Camera should be reoriented  $\leftarrow$ 
11:   else
12:    Camera should be reoriented  $\rightarrow$ 
13:   end if
14: Reorient the robot camera  $\Theta = \Theta/2$  along required direction { $\Theta$  is same in all the directions for NAO robot}
15: Recapture a new photograph
16: Calculate  $f_{th}$ ,  $f_{gr}$ 
17: end while

```

In this approach, the aesthetic features of the recaptured image at every stage are compared to corresponding thresholds at every stage. Figure 7 shows an example with intermediate stage taken by the NAO robot. For a given angle view range of the robot camera, the number of photographs taken is bounded by $\lceil \log_2(angle_view_range) \rceil - 1$. In our experiments the maximum number of photos taken were six.

3.3 Discussion

The change in robot reorientation can also be determined by computing the depth of the focused subject and later using properties of similar triangles. It can be accomplished using depth estimation algorithms from computer vision [Torralba and Oliva, 2003; Saxena *et al.*, 2005]. These approaches are computationally intensive, with [Saxena *et al.*, 2005] taking 78 seconds to compute the depth map in Figure 8. The closer regions are represented in yellow and the farther regions in blue.

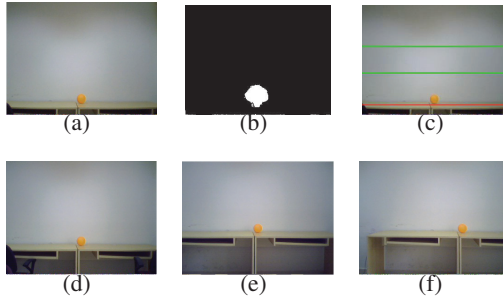


Figure 7: Important intermediate stages of robot reorientation; (a) Initial photograph, (b) Extracted subject region, (c) Detected horizon line in red, (d) Reorienting robot camera by 8° ↓, (e) Reorienting robot camera by 4° ↓, (f) Final image obtained after reorienting the robot camera towards left by 2°

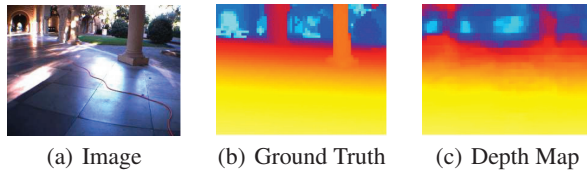


Figure 8: Depth map generation

4 Results

4.1 Image Dataset

We first demonstrate the performance of our image quality assessment approach on a large diversified photo database collected by [Ke *et al.*, 2006]. The database was acquired by crawling a photography contest website, DPChallenge.com, which contain a large number of images taken by different photographers. These images are rated according to their visual appeal by the photographers community. The average of the rating values of an image is used as ground truth to classify them into high and low quality categories. Out of the obtained 60000 ranked images the top 6000 images are chosen as the high quality images and the bottom 6000 as the low quality images. Of the 6000 images in each category, randomly selected 3000 images are used for training and the other 3000 images for testing.

We achieved an accuracy of 79% on [Ke *et al.*, 2006] database using a two class SVM classifier. Extracting all the high level and the aesthetic features of an image took approximately 2 seconds by our approach compared to a minimum of 14 seconds of our best possible implementation of [Luo and Tang, 2008] in matlab. The 2 seconds time taken by our approach is the maximum time taken by an image from the 12,000 images (6,000 good, 6,000 bad). Some of the 21% error in the accuracy can be accounted to the photographs that either follow the other guidelines of photography like the diagonal rule etc., or those which do not follow any of the guidelines. The performance can be improved by increasing the number of features and a more sophisticated design of these statistical high level parameters of an image. Also [Luo and Tang, 2008] could achieve the 93% success rate be-

Table 2: Results of saliency based quality assessment on photographs from Ke *et al.* dataset

Input Image	Subject Region	GroundAssessedDeviation		
		Truth	Quality	(f_{th}, f_{gr})
		Good	Good	(0.07,0.11)
		Bad	Bad	(0.14,0.07)
		Good	Bad	(0.17,0.45)
		Bad	Good	(0.19,0.42)

cause of the complex computations which help in extracting the subject region with much accuracy.

Despite the fact that we are choosing the top 10% and the bottom 10% of the 60,000 images, there is significant overlap in the individual rating distribution. The class separability between the good and bad images improves if we restrict ourselves to the top and bottom 2% of the 60,000 images. As the individual rating values of Ke's dataset were not available we collected another dataset of 60,000 images from DPChallenge.com. When the class separability is high (top/bottom 2%) there are no false positives but with the top/bottom 10% there were false positives of about 7%. It is observed that with less class separability, the percentage of false positives increase. To reduce the false positives a more sophisticated solution is required. Table 2 show the results on few images. Table 3 shows the results of experiments on where we tested on top and bottom 2-10% keeping the training set constant on our dataset and table 4 shows the comparison of results on Ke's dataset.

Table 3: Testing on top and bottom n% of our dataset

	10%	8%	6%	4%	2%
Error rate	21%	20%	18%	15%	11%

Table 4: Comparison of performance on Ke's dataset

	Ke et al. 2006	Luo et al. 2008	Our Approach
Accuracy	72%	93%	79%

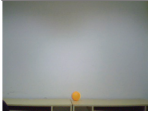
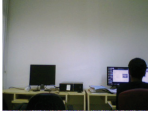
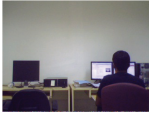
4.2 NAO Robot

We test our system on the humanoid NAO robot [Gouaillier *et al.*, 2008]. It has two fixed focus cameras, one in the forehead region and other at the chin region which do not form a stereo pair. It has a fixed aperture size and shutter speed. The NAO can rotate its head in all four directions, up, down

$[-119.5^\circ, 119.5^\circ]$; and left, right $[-38.5^\circ, 29.5^\circ]$. The angle view range of the NAO's camera is 34° .

We trained the robot using all the 6000 good and 6000 bad images from the [Ke *et al.*, 2006] dataset. In our experiments, we perform the robot reorientation methodology on the $(\Theta =) 16^\circ$ view of the camera. Table 5 presents few results of our approach on NAO. Our results show that robots can be programmed to capture better photographs.

Table 5: Performance of our approach on NAO with last column showing the number of images recaptured

Initial image	Intermediate directions, (f_{th}, f_{gr})	Final Photo	(f_{th}, f_{gr})
	$8^\circ \uparrow, 4^\circ \uparrow, 8^\circ \rightarrow, 4^\circ \leftarrow$ (0.30, 8.76)		(0.04, 0.41)
	$8^\circ \downarrow, 4^\circ \uparrow, 2^\circ \uparrow, 8^\circ \rightarrow, 4^\circ \leftarrow, 2^\circ \leftarrow$ (0.28, 1.75)		(0.09, 0.18)
	$8^\circ \downarrow, 4^\circ \uparrow$ (0.33, 2.02)		(0.11, 0.04)

The second row of Table 5 shows an example with enhanced visual appeal. The last experiment in the third row shows a part of the ball being occluded initially, which when recaptured is a better image that is preferable for processing. This makes us believe that aesthetic quality can aid processing of images.

5 Conclusion

This research helps a robot to recapture a better photograph (if required) by assessing the visual quality of the captured photo. The strength of our approach is the computational efficiency which can be applied in autonomous robots. The accuracy can be improved further by adding symmetry in the subject region as mandatory since images with some symmetry are rated higher than the rest and with more complicated composition guidelines of professional photography. We believe that with some changes to the pose of the robot we can get better visually appealing images. One direction of our future work is focused on accurately estimating the desired change in the pose of the robot for taking better photographs. For the next version of our system, we will use a robot camera which supports manual focus, manual exposure (by adjusting aperture value and shutter speed), and much higher resolution.

References

[Achanta *et al.*, 2009] R Achanta, S Hemami, F Estrada, and S Susstrunk. Frequency-tuned salient region detection, 2009.

[Ahn *et al.*, 2006] H Ahn, D Kim, J Lee, S Chi, K Kim, J Kim, M Hahn, and H Kim. A robot photographer with user interactivity. In *IROS*, 2006.

[Ang, 2004] T Ang. *Digital Photographer's Handbook*. 2004. Dorling Kindersley Limited, UK.

[Byers *et al.*, 2003] Z Byers, M Dixon, K Goodier, C M Grimm, and W D Smart. An autonomous robot photographer. In *IROS*, pages 2636–2641, 2003.

[Gouaillier *et al.*, 2008] D Gouaillier, V Hugel, P Blazevic, C Kilner, J Monceaux, P Lafourcade, B Marnier, J Serre, and B Maisonnier. The nao humanoid: A combination of performance and affordability. In *IEEE Transactions on Robotics*, 2008.

[Grill and Scanlon, 1990] T Grill and M Scanlon. *Photographic Composition*. 1990. American Photographic Book Publishing.

[Harris, 2010] Dan Harris. What make a good photograph. In *High Definition Professional Photography*, 2010. <http://www.danharrisphotoart.com>.

[Hou and Zhang, 2007] X Hou and L Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[Ke *et al.*, 2006] Y Ke, X Tang, and F Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[Kim *et al.*, 2010] M Kim, T Song, S Jin, S Jung, G Go, K Kwon, and J Jeon. Automatically available photographer robot for controlling composition and taking pictures. In *IROS*, 2010.

[Lamb and Stevens, 2010] J Lamb and R Stevens. The eye of the photographer. In *The Social Studies Texan*, volume 26, pages 59–63, 2010.

[Leykin, 2006] Alex Leykin. Vanishing points. 2006. <http://www.cs.indiana.edu/~sjohnson/irrt/src/index.htm>.

[Luo and Tang, 2008] Y Luo and X Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008.

[Ma and Zhang, 2008] Q Ma and L Zhang. Image quality assessment with visual attention. In *ICPR*, 2008.

[Saxena *et al.*, 2005] A Saxena, S H Chung, and A Y Ng. Learning depth from single monocular images. In *NIPS*, 2005.

[Tong *et al.*, 2004] H Tong, M Li, H Zhang, J He, and C Zhang. Classification of digital photos taken by photographers or home users. In *Pacific Rim Conference on Multimedia*, pages 198–205. Springer, 2004.

[Torralba and Oliva, 2003] A Torralba and A Oliva. Depth estimation from image structure. In *PAMI*, volume 24, pages 1226–1238, 2003.

[Wang *et al.*, 2002] Z Wang, H R Sheikh, and A C Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *ICIP*, 2002.