Single Image Depth Level Estimation Using Dark Channel Prior

Yichao Cao\textsuperscript{1,2} and Xiaobo Lu\textsuperscript{1,2,\ast}

\textsuperscript{1} School of Automation, Southeast University, Nanjing 210096, China
\textsuperscript{2} Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China

\ast E-mail: xblu2013@126.com

Abstract. In this paper, we present a simple image depth level estimation algorithm. From the dark channel prior theory, an estimate of the air transmittance in the image is calculated. In wild surveillance, the disparity in the image poses a huge challenge for smoke detection and other video analysis tasks. Appropriate depth level estimation provide significant prior knowledge for subsequent identification and detection. For landscape images, we can approximate the air transmittance to depth information for histogram analysis. The depth value is segmented by a multi-threshold segmentation algorithm, and the resulting image can be used for forest fireworks detection and the like. This method does not rely on samples and classifiers, and the algorithm does not require training. The final experimental results show that the depth level estimation of a single landscape image based on the dark channel prior can achieve good results.

1. Introduction

In optics, the optical phenomenon known as depth of field, is the distance about the Plane of Focus where objects appear acceptably sharp in an image. In some cases, it may be desirable to have the entire image sharp, and a large depth of field is appropriate. Depth of field is an important geometric relations within images taken by camera, and is of great significance in image processing and video analysis. In real life, even in sunny days, there are some particles in the air, so people can have a deep concept when observing the scene.

In applications such as forest fire prevention, depth estimation has an essential impact on algorithm performance. If the algorithm parameters of different depth of field are the same, it is bound to cause two situations: the shaking of trees near the field of view is easy to trigger false detection because the pixel area is large, and the distant smoke is easy to miss due to the small pixels occupying the image. The depth of field estimate for this scene is different from the general depth of field estimate. Instead of calculating a particularly accurate depth of field image, it estimates the region of interest through a relative depth map.

While there is much prior work on estimating depth based on stereo images or motion \cite{1}, there has been relatively little on estimating depth from a single image. Most work on estimating depth has focused on using methods such as stereovision or structure from motion \cite{2}, which require two (or more) images.

Obtaining accurate depth of field estimates from a single image is undoubtedly very difficult. In video content analysis, appropriate depth level estimation provide significant prior knowledge for...
The method proposed in this paper is shown in Figure 1. First, the dark channel image of the source image is calculated. Secondly, the air transmission is calculated according to the atmospheric light and image forming model. Finally, the air transmission is threshold-divided to obtain the hierarchical depth estimation.

**2. Related Work**

Directly related to our work are several methods of estimating depth from a single image. Saxena et al. [3] use a Markov Random Field (MRF) to infer a set of “plane parameters” that capture both the 3-d location and 3-d orientation of the patch. However, the method for inferring detailed 3-d structure from a single still image rely on small homogeneous regions in the image and suffers in less controlled settings. Hoiem et al. [4] labels regions of the input image into coarse categories: “ground”, “sky”, and “vertical” which is similar to our work. These labels are then used to “cut and fold” the image into a pop-up model using a set of simple assumptions. But this method is not expected to work on our wide scenes, their boundaries are not clear.

Ladicky et al. [5] proposed a new pixel-wise classifier that can jointly predict a semantic class and a depth label from a single image. The main weaknesses of the method are the inability to deal with low resolution images, very large requirements in terms of hardware and apparent inability to locate the objects more precisely for semantic classes with high variance.

Konda et al. [6] presented a deep learning approach for joint estimation of disparity and motion based on learning about interrelations between images from multiple cameras, multiple frames in a video, or the combination of both.

**3. Dark Channel Prior**

The dark channel prior theory is based on statistics on a large number of outdoor pictures. The study found that [7], in most non-sky local areas, some pixels always have at least one color channel with a very low value. In other words, the minimum value of the light intensity in the area in the picture is a small number.

**3.1. The dark channel prior theory**

For any input image $J$, its dark channel can be expressed as:

$$J^\text{dark}(x) = \min(\min(J^c(y)))_{c\in\{r,g,b\}, y\in\mathcal{D}(x)}$$

(1)
Where \( j^c(x) \) is a color channel of input image \( J \), and \( \Omega(x) \) is a local patch centered at \( x \). In the actual calculation, we first find the minimum value of the three channels of each pixel, store it in a grayscale image with the same size as the original image, and then filter the grayscale image with minimum value. The radius is determined by the size of the window, generally with WindowSize = 2 * Radius + 1.

The dark channel prior theory states that \( j_{\text{dark}}(x) \) should be a small number close to zero. The pixel value of the dark channel of the near image in the actual scene is very low. We compared some outdoor forest surveillance images. As shown in Figure 2, the dark channel values of the image pixels at different distances in the picture differ greatly. The main idea of this paper is to make a deep estimate by analyzing this difference.

### 3.2. Estimating the transmission

In computer vision and computer graphics, the graph formation model described by the following equations is widely used:

\[
I(x) = J(x)t(x) + A(1 - t(x)) \tag{2}
\]

Where \( I(x) \) is the image captured by the camera, \( J(x) \) is the ideal image, \( A \) is the global atmospheric light component, and \( t(x) \) is the transmission. The known condition is \( I(X) \), the purpose is to calculate \( J(x) \). Obviously, this is an equation with infinite solutions, so some priori constraints are needed.

**Figure 2.** The dark channel image uses a window size of 15×15 and a minimum filter radius of 7 pixels during the calculation.

In our approach, we select the top 0.1% brightest pixels in the dark channel which are the most opaque. Among these pixels, the highest intensity pixel mean in the input image is selected as atmospheric light \( A \). Assume that the transmission \( t(x) \) is constant in each window. According to the dark channel prior theory mentioned above, the estimated value of transmission can be derived:

\[
\hat{t}(x) = 1 - \min_{y \in \Omega(x)} \left( \frac{j_{\text{dark}}^c(y)}{A} \right) \tag{3}
\]
In practice, even on sunny days, the atmosphere is not absolutely free of any particles. Therefore, when we use camera to capture images, the haze in the image still exists. In addition, the existence of haze is the basis of human perception of the depth of the scene. This phenomenon is called aerial perspective [8, 9], which also is the basis for our deep estimation method.

As shown in figure 3, the calculated air transmission map is approximately similar to the true deep value of the image. But note that what we get is relative image depth, not an absolute depth value. In order to adapt to different scenarios and obtain robust depth estimation, we need to do multi-threshold segmentation on the obtained relative image depth.

![Image](image_url)

**Figure 3.** The calculated air transmission map

### 4. Multi-threshold Segmentation

In the calculation of the air transmission, the image we get is a grayscale image. In order to further segment the image, we need to use multi-threshold segmentation technology.

#### 4.1. Single threshold segmentation

Suppose the grayscale space of an image is represented by L. Use i for each grayscale value in the grayscale space. The number of occurrences of each grayscale value in the grayscale histogram in the image is denoted by \( n_i \) and the total number of pixels of one image is denoted by \( N \), then \( N = n_1 + n_2 + \ldots + n_L \). For each grayscale value, the probability of appearing on the entire image is \( f_i \).

If a single threshold segmentation is employed, the grayscale set is divided into \( G_0 \) and \( G_1 \). And \( G_0 = \{0,1,2, \ldots, t\}, G_1 = \{t + 1, t + 2, \ldots, L - 1\} \). The total mean value of the image is \( \mu_T \), and the mean values of the regions corresponding to \( G_0 \) and \( G_1 \) in the image are \( \mu_0 \) and \( \mu_1 \) respectively:

\[
\mu_0 = \sum_{i=0}^{t} \frac{if_i}{\omega_0} \quad (4)
\]

\[
\mu_1 = \sum_{i=t+1}^{L-1} \frac{if_i}{\omega_1} \quad (5)
\]

\[
\mu_T = \sum_{i=0}^{t-1} if_i + \omega_0 \mu_0 + \omega_1 \mu_1 \quad (6)
\]
Where \( \omega_0 \) and \( \omega_1 \) are the probability ratios of the two types of pixels G1 and G2. The inter-class variance \( \sigma_B^2 \) and the intra-class variance \( \sigma_W^2 \) of the two regions are respectively:

\[
\sigma_B^2 = \omega_0 (\mu_0 - \mu_T)^2 + \omega_1 (\mu_1 - \mu_T)^2
\]

\[
\sigma_W^2 = \omega_0 \sigma_0^2 + \omega_1 \sigma_1^2
\]

When using the OTSU method [10] to segment an image, the optimal threshold \( t \) maximizes the inter-class variance or minimizes the intra-class variance. This threshold \( t \) is the segmentation threshold we are looking for. Because in the OTSU method it can be inferred that the relationship between the inter-class variance and the intra-class variance is inversely proportional. Another way to use the maximum inter-class variance to determine the optimal threshold \( t_{opt} \) is to use the smallest intra-class variance for optimal discrimination:

\[
t_{opt} = \underset{t \in [\text{thresholds}]}{\text{Max}} \{\sigma_B^2(t)\} = \underset{t \in [\text{thresholds}]}{\text{Min}} \{\sigma_W^2(t)\}
\]

The principle of maximum inter-class variance is consistent with the principle of minimum intra-class variance, but because the maximum inter-class variance is relatively simple in calculation, the principle of maximum inter-class variance is usually used to calculate the optimal threshold.

### 4.2. Multi-threshold segmentation

It is natural to extend the OTSU method to multi-threshold, that is, the multi-threshold OTSU algorithm. Use \( i \) for each grayscale value in the grayscale space. The number of occurrences of each grayscale value in the grayscale histogram in the image is denoted by \( n_i \) and the total number of pixels of one image is denoted by \( N \), then \( N = n_1 + n_2 + \ldots + n_L \). For each grayscale value, the probability of appearing on the entire image is \( \frac{n_i}{N} \).

For the multi-threshold segmentation algorithm, the image needs to be segmented into \( n \) classes, and it is obvious that \( n-1 \) thresholds are needed to segment the image. Assume that \( n-1 \) thresholds are \( t_1, t_2, \ldots, t_{n-1} \). The grayscale set can be divided into \( G_0 = \{0,1,2,\ldots, t_1\}, G_1 = \{t_1 + 1, t_1 + 2, \ldots, t_2\}, \ldots, G_{n-1} = \{t_{n-1} + 1, t_{n-1} + 2, \ldots, L - 1\} \). The total mean value of the image is \( \mu_T \), the mean values of the regions corresponding to \( n \) classes in the image are \( \mu_0, \mu_1, \ldots, \mu_{n-1} \), and the probability of occurrence corresponding to \( n \) classes is \( F_0, F_2, \ldots, F_{n-1} \).

Similar to the single threshold segmentation method, the optimal threshold for multi-threshold segmentation algorithm needs to follow the principle of maximum inter-class variance or minimum intra-class variance. The optimal thresholds can maximize the inter-class variance:

\[
\text{arg}_{t} \max (\sigma_B^2) = \sum_{i=0}^{n-1} F_i (\mu_i - \mu)^2
\]

### 5. Experimental Results

In our experiments, the patch size is set to \( 15 \times 15 \) for a \( 1400 \times 800 \) image. It takes about 0.4 seconds to process a \( 1400 \times 800 \) pixel image on a PC with a 2.2 GHz Intel Core i5-5200U Processor.

The depth level of field estimate for this scene is different from the general depth of field estimate. One advantage of this estimation algorithm is that the execution speed is fast, and in a practical application, it is necessary to calculate only once for a fixed scene. Another advantage of this method is that it does not rely on samples and classifiers, and the algorithm does not require training.

As shown in Figure 4, (a) is the input image and (b) is the calculated atmospheric transmission map. The range of variation is different for different lighting and scenes. And (c) is a double-threshold segmentation picture. After the double threshold segmentation, the picture can be basically divided into three parts: sky, distant view and close view. This basically satisfies the general visual recognition needs, such as forest fireworks detection and other tasks.
6. Discussions and Conclusions
In this paper, we present a simple image depth level estimation algorithm from single image. This method does not rely on samples and classifiers, and the algorithm does not require training. Multi-threshold segmentation image can provide significant prior knowledge for subsequent identification and detection.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No.61374194), the National Key Science & Technology Pillar Program of China (No.2014BAG01B03), Key Research and Development Program in Jiangsu Province (No.BE2016739) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References
[1] Scharstein D, Szeliski R, Zabih R. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms[C]// Stereo and Multi-Baseline Vision. IEEE, 2002:7-42.
[2] D. A. Forsyth and J. Ponce. Computer Vision: A Modern Approach. Prentice Hall, 2003.
[3] Saxena A, Sun M, Ng A Y. Learning 3-D Scene Structure from a Single Still Image[C]// IEEE, International Conference on Computer Vision. IEEE, 2007:1-8.
[4] Hoiem D, Efros A A, Hebert M. Automatic photo pop-up [J]. Acm Transactions on Graphics, 2005, 24(3):577-584.
[5] Shi J, Pollefeys M. Pulling Things out of Perspective[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014:89-96.
[6] Konda K, Memisevic R. Unsupervised learning of depth and motion [J]. Computer Science, 2013.
[7] He K, Sun J, Tang X. Single Image Haze Removal Using Dark Channel Prior [J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(12):2341-2353.
[8] James M K, James K, Goldstein E B. Sensation and Perception [J]. Leonardo, 1983, 15(2):160.
[9] Preetham A J, Shirley P, Smits B. A practical analytic model for daylight[C]// Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co. 1999:91-100.
[10] Otsu N. A Threshold Selection Method from Gray-Level Histograms [J]. IEEE Trans.syst.man. & Cybert, 2007, 9(1):62-66.