Segmenting focused objects based on the Amplitude Decomposition Model

Tiantang Chen, Hongliang Li

School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract

In this paper, we present an inherent model of the low depth-of-field (DOF) images, referred as the Amplitude Decomposition Model, which turns out to be useful for the detection and segmentation of focused objects in the low DOF images. By analyzing the low DOF image in frequency domain, the Amplitude Decomposition Model is firstly investigated, i.e., the amplitude spectrum of the low DOF image can be decomposed into the amplitude of its totally defocused version and the high-frequency difference amplitude of its focused regions. Based on this model, we propose a method for detecting focused objects. Using the detection result, we then utilize a thresholding method to segment the focused objects and employ the graph cut technique to refine the focused object boundary. Experimental results show that the proposed method can extract focused objects effectively and is comparable to the state-of-the-art methods.

1. Introduction

As the convenient access of digital equipments, photos and videos are emerging in our daily life. Among them, a kind of image called the low depth-of-field (DOF) image is favored because of its unique visual effect. Taken at the proper focus position of the camera lens, these images usually have the object of interest in sharp focus, while the background is blurred. The focused object preserves the details of the scene, and turns out to be dominated by high frequency components. On the other hand, the defocused background contains significant low frequency components. The high contrast in the low DOF images helps people pay more attention to the focused object and make the low DOF images prevail in TV programs or film productions.

In general, focused objects correspond to the semantic objects which represent the visual information and semantic content of the images or videos. The semantic object usually plays an important role in many applications such as adaptive object-based image compression (Pan, 2002), object recognition (Hu, 2011) and video surveillance (Gurwicz et al., 2011). However, semantic object segmentation (Mailing and Frlas, 2011) is still a challenging task in computer vision and pattern recognition. Generally, object segmentation can be completed by two steps, i.e., object detection and object region extraction. The first step aims to train a classifier from a lot of samples. The second step is to partition the image into foreground and background regions. Traditional segmentation approaches include region splitting or merging (Pavlidis and Liow, 1990), active contours (Blake and Isard, 1998), mean shift (Comaniciu and Meer, 2002), etc. One of the notable developments in the last decade is the graph based solution. Boykov and Jolly (2001) formulated the segmentation problem as the minimization of an energy function on a graph, performed the optimization using the minimum cut procedure. Shi and Malik (2000) normalized the volume of each segment, simplified the minimum cut problem into a generalized eigenvalue problem. Furthermore, Grady (2006) considered segmentation as a graph-based random walk problem, assigned each pixel with class labels according to the probability that a random walker first reaches the seed point.

The work of segmenting the low DOF images can be traced back to the paper in 1990s (Tsai and Wang, 1998). The approach is based on the observation that there is small amount of defocus in the edges of focused objects while large amount of defocus in the edges of background objects. The amount of defocus is then estimated over each edge pixel using the moment-preserving method. Finally, the closed boundary of focused objects is obtained by an edge-linking procedure.

Unlike the edge-based approaches, a multiscale approach based on high frequency wavelet coefficients was presented by Wang et al. (2001). The method does not rely on the process of connecting object boundaries. In contrast, it is motivated by the observation that the focused regions have more high-value wavelet coefficients in the high frequency bands of the wavelet transform. By analyzing the details of the regions in terms of the statistics of the coefficients, the method is more robust than the edge-based segmentation algorithms.

On the other hand, the method proposed by Kim (2005) depends on high-frequency contents. By computing higher order statistics (HOS) for each pixel, the low DOF image is transformed into a feature space called the HOS map. To eliminate errors and noises,
a morphological filter is then used to simplify the HOS map. Finally, region merging and adaptive thresholding are employed to extract the focused object.

In the paper of Li and Ngan (2007), the authors proposed an unsupervised segmentation algorithm which models the low DOF image as a matting problem. A reblluring model is built to generate the focus map of the input image, which indicates the location of the focused object. Next, bilateral and morphological filtering are applied to smooth and merge the focused regions. Then, an adaptive error control matting approach is used to perform the final segmentation and boundary refinement.

In the recent work of Liu et al. (2010), a focus energy map is first generated based on the difference of high-frequency components between focused region and defocused background. Then, the focus energy map is used to construct region/boundary saliency maps. Next, a boundary linking method is employed to obtain closed region/boundary masks, which are exploited to generate a trimap. Finally, image matting is performed on the trimap to segment focused objects.

In this paper, we propose a new method to extract focused objects from low DOF images. This method consists of three steps. The first step is to generate the focus maps based on the Amplitude Decomposition Model. The focus maps can achieve the goal of focused object detection, identifying focused objects effectively. The second step is to segment the low DOF images by thresholding over the focus maps. This step can provide the initial segmentations of the low DOF images. In the third step, the graph cut technique is employed to refine the segmentation results.

The rest of the paper is organized as follows. In Section 2, we analyze the low DOF images in frequency domain to reveal the Amplitude Decomposition Model. A detection method for low DOF images is then proposed, and a scheme for focused object extraction is presented. In Section 3, experimental results are provided to evaluate the performance of the method. Finally, Section 4 draws the conclusion.

2. The proposed method

2.1. The Amplitude Decomposition Model

As mentioned above, a low DOF image has proper focused objects but a blurry background. In contrast, a defocused image refers to the one that is totally out-of-focus, i.e., both the objects and the background are blurred. Fig. 1 shows some example images, which include a clear image, a low DOF image and a defocused image.

According to Li and Ngan (2007), a low DOF image can be considered as the composite of a clear image and a defocused image using a matting model:

\[
i(x, y) = \alpha(x, y)i_c(x, y) + (1 - \alpha(x, y))i_d(x, y),
\]

where \(i_c(x, y)\) denotes the low DOF image in RGB color space, \(i_c(x, y)\) denotes the clear image and \(i_d(x, y)\) denotes the defocused image, with \((x, y)\) being the coordinates of the images. And \(\alpha(x, y)\) is a binary function, equaling 1 for the focused regions while 0 for the defocused regions, i.e.

\[
\alpha(x, y) = \begin{cases} 
1, & \text{if } (x, y) \in \Omega_1; \\
0, & \text{if } (x, y) \in \Omega_0, 
\end{cases}
\]

Here \(\Omega_1\) and \(\Omega_0\) denote the focused and defocused regions, respectively.

Using the Fourier transform, we can rewrite Eq. (1) in the frequency domain,

\[
\hat{i}(u, v) = \hat{\alpha}(u, v) \hat{i}_c(u, v) + |\hat{\alpha}(u, v) - \hat{\alpha}(u, v)| \hat{i}_d(u, v),
\]

where the symbols with hat are the Fourier transforms of the original signals, \((u, v)\) is spatial frequency, and \(*\) denotes the convolution operation.

For simplicity, we transfer it to a polar coordinate with \(u = f \sin \theta\) and \(v = f \cos \theta\). Average over the angle by computing spatial frequency \(f = \sqrt{u^2 + v^2}\), we can express Eq. (3) in one dimension:

\[
\hat{i}(f) = \hat{\alpha}(f) \hat{i}_c(f) + |\hat{\alpha}(f) - \hat{\alpha}(f)| \hat{i}_d(f),
\]

where

\[
\hat{i}(f) = F(i(x, y)), \quad \hat{\alpha}(f) = F(\alpha(x, y)), \quad \hat{i}_c(f) = F(i_c(x, y)), \quad \hat{i}_d(f) = F(i_d(x, y)).
\]

Let us focus on the amplitude spectrum, it is given by

\[
|\hat{i}(f)| = |\hat{i}_c(f)| + |\hat{\alpha}(f) - \hat{\alpha}(f)| \left(1 - A e^{\frac{\lambda^2}{2}}\right).
\]

Previous studies indicated that the amplitude spectrum of natural images behaves approximately as \(f^{-2}\) (Field and Brady, 1997) \(\lambda\) ranges from 0.6 to 1.6. A clear image has a relatively uniform frequency distribution, which results in a small \(\lambda\). Meanwhile, a defocused image can be viewed as the convolution of a Gaussian and the

\[\]
clear version of the same scene, which leads to a large $i$, because of the reduced amplitude in high frequency after the Gaussian blurring. Fig. 2 plots the amplitude spectrum of the example images displayed in Fig. 1. For a better view, the amplitude is drawn on a log scale. We can see that the amplitude spectrums of the low DOF image and the defocused image are steep than that of the clear image. Here we use the high-frequency difference amplitude $d(f)$ to represent the high-frequency amplitude loss of the defocused image with respect to the clear image.

Based on the observation above, Wiener (1958) attempted to restore the high frequency parts of the defocused images to the original amplitude of their clear versions to solve the “deblurring” problem. In contrast, we perform in another direction—we decompose the amplitude of the clear image into two components, i.e., the amplitude of its defocused version and the high-frequency difference amplitude, given by

$$
|\hat{i}(f)| = |\hat{u}(f)| + d(f) = |\hat{u}(f)| \cdot Ae^{\frac{p^2}{2}} + d(f).
$$

Therefore, we have

$$
1 - Ae^{\frac{p^2}{2}} = \frac{d(f)}{|\hat{u}(f)|}.
$$

Then Eq. (8) can be written as

$$
|\hat{i}(f)| = |\hat{u}(f)| + \frac{|\hat{z}(f) \ast \hat{i}(f)| \cdot d(f)}{|\hat{i}(f)|} = |\hat{u}(f)| + \frac{|\hat{z}(f) \ast \hat{i}(f)|}{|\hat{i}(f)|} \cdot d(f) = |\hat{u}(f)| + d_i(f)
$$

with

$$
d_i(f) = \frac{|\hat{z}(f) \ast \hat{i}(f)|}{|\hat{i}(f)|} \cdot d(f).
$$

Therefore, from Eq. (11), we can see that the amplitude spectrum of the low DOF image has been decomposed into two terms. The first term corresponds to the amplitude of its defocused version. As for the second term, we can infer that it corresponds to the high-frequency difference amplitude of the focused regions. We simply call this as the Amplitude Decomposition Model (ADM), which provides an obvious cue for developing a method of detecting focused objects from low DOF images. Specifically, we subtract the amplitude of the defocused version from that of the low DOF image, then the high-frequency difference amplitude corresponding to the focused objects is obtained. Finally, we construct the focus map by transforming the difference amplitude into the spatial domain.

However, we only have the low DOF image in hand, its clear version is unknown. It is difficult to estimate the amplitude of the clear version $|\hat{u}(f)| = f^{-1}$, not to mention that the value of $\lambda$ varies from different images. Therefore, we can not obtain the amplitude of the defocused version $|\hat{u}(f)|$ by Eq. (6). Here we attempt to approximate $|\hat{u}(f)|$ by averaging $|\hat{i}(f)|$ with an average filter $h_n(f)$, i.e.

$$
|\hat{u}(f)| = |\hat{i}(f)| \ast h_n(f).
$$

(13)

$$
h_n(f) = \frac{1}{n} \{1 \cdots 1\}, n.
$$

(14)

Note that $h_n(f)$ is an $n \times n$ matrix when applied over two-dimension signals. According to our empirical study, good performance can be achieved when $n$ takes value between 300 and 500. In our experiments described in the following sections, we adopt the average filter with $n = 300$.

Based on the Amplitude Decomposition Model, we now present the method for detecting focused objects from low DOF images in Algorithm 1.

**Algorithm:** Focused object detection

**Input:** A low DOF image $(x,y)$

**Output:** The focus map $FM(x,y)$

1. Find the amplitude spectrum $|\hat{i}(f)|$ and the phase spectrum $p(f)$ by

$$
|\hat{i}(f)| = |F(i(x,y))|
$$

$$
p(f) = \exp(i \cdot p(f))
$$

Here $F$ denotes the Fourier transform

2. Find the high-frequency difference amplitude of focused regions $d_i(f)$ by

$$
d_i(f) = |\hat{i}(f)| - |\hat{i}(f)| \ast h_n(f)
$$

Here $h_n(f)$ is an average filter with $n = 300$

3. Define the focus map of the low DOF image as

$$
FM(x,y) = g(x,y) \ast \left| F^{-1}[d_i(f) \cdot \exp(i \cdot p(f))]|^2
$$

Here a Gaussian filter $g(x,y)$ is applied to eliminate the noise effect (the variance of the Gaussian filter is set to 2.5 according to our empirical study) and $F^{-1}$ denotes the inverse Fourier transform

2.2. Focused object extraction

Using the focus detection result, we now turn to extract focused objects from the low DOF images. In the method, we first employ the proposed detection method to generate the focus maps of the low DOF images. Examples of the focus maps can be found in Fig. 3(b) and Fig. 4(c). We can see that the white region in the focus maps indicates the focused objects and provides an effective reference for the possible location of focused objects.

Next, we binarize the focus map to get the initial segmentation result. The simplest method to do this is thresholding over the focus map. In our implementation, the mean value of the focus map
is chosen as the threshold. The pixels whose focus value is greater than the threshold are classified as the focused objects, otherwise the background:

\[(x, y) \in \{ \Omega_f, \text{ if } FM(x, y) \geq \text{mean}(FM); \Omega_b, \text{ otherwise,} \} \quad (15)\]

here 'mean' denotes the mean operation. When the segmentation is obtained, morphological operations are performed to fix the errors.

Finally, we employ GrabCut (Rother et al., 2004) to refine the segmentation result. The main idea of GrabCut is iteratively using graph cut (Boykov and Jolly, 2001) and Gaussian mixture models (GMM). Specifically, we first perform erosion and dilation operations on the initial focused object boundary to generate a trimap which consists of three regions, i.e., the foreground, the background and the unknown. In the implementation, we erode 3 pixels and dilate 15 pixels on the either side of the current focused object boundary. Then we utilize graph cut to assign labels to the pixels in the unknown region, using the implementation of Bagon (2006). Next, we introduce morphological operations to fill the holes in the segmentation result. The iteration repeats at most 8 times until the labels in the segmentation result keep unchanged.

As for graph cut, it is an optimization method defined on a graph \( G = (V, E) \), here \( V \) is a set of nodes and \( E \) is a set of undirected edges which connect the nodes. The optimization can be achieved by solving the min-cut/max-flow problem, which is accomplished by solving an energy function:
The data cost $E_i$ encodes the cost when node $i$ is assigned the foreground/background label $x_i$. It is measured by the distance between node $i$ and the foreground/background terminal node. In order to model the foreground/background color distributions, Gaussian mixture models with 5 and 3 components are used to approximate the regional properties of the foreground and the background, respectively. The mean and covariance of each component are estimated using K-means. For a pixel $i$ with the color $z_i$ and the label $x_i$, its distances to the foreground and the background are defined as follows:

$$E_i(x_i = 1) = \min_k \left( d^f_k, k \in [1, 5] \right),$$

$$E_i(x_i = 0) = \min_k \left( d^b_k, k \in [1, 3] \right)$$

with

$$d^f_k = -\log \left( w_k \left( \frac{1}{\det \Sigma_k} \right)^{1/2} \exp \left( -\frac{1}{2} \left( z_i - \mu_k \right)^T \Sigma_k^{-1} \left( z_i - \mu_k \right) \right) \right),$$

$$d^b_k = -\log \left( w_k \left( \frac{1}{\det \Sigma_k} \right)^{1/2} \exp \left( -\frac{1}{2} \left( z_i - \mu_k \right)^T \Sigma_k^{-1} \left( z_i - \mu_k \right) \right) \right),$$

where $w_k$ denotes the weights for the $k$th component of the GMM, and $\mu_k$ and $\Sigma_k$ are the means and the covariances, respectively.

The smoothness cost $E_2$ encodes the cost when two adjacent nodes $i$ and $j$ are assigned different labels. It can be measured by the similarity between pixels $i$ and $j$, defined as:

$$E_2(x_i, x_j) = \| x_i - x_j \| \cdot \exp \left( -\beta \| z_i - z_j \|^2 \right),$$

where $\| z_i - z_j \|$ is the $L_2$ norm of the RGB color difference, and $\beta$ is a robust parameter that weights the color contrast (Li et al., 2005).

Algorithm 2 summarizes the segmentation process and Fig. 3 shows the example results obtained in each step. Note that if there are multiple focused objects in one image, we generate the trimap based on the location of the focused object with the largest area. For the trimap illustrated in Fig. 3(d), the green region is the foreground which provides the data for the new foreground model (GMM) in the next iteration, while the blue region is the background which provides the data for the new background model. And the region left unchanged is the unknown region where the next classification is performed.

**Algorithm 2.** Focused object extraction

**Input:** A low DOF image  
**Output:** Pixel labels  
1: Generate the focus map based on the ADM  
2: Binarize the focus map by thresholding  
3: Fix the errors by the morphological operations  
4: Generate the trimap by the erosion and dilation operations  
5: Classify the pixels in the unknown region by graph cut  
6: Fix the errors by the morphological operations  
7: Iterate from Step (4) until the labels in the segmentation result keep unchanged

### 3. Experiments

In this section, we evaluate our method on several low DOF images, and do some comparisons with two methods: Kim’s method (Kim, 2005) and Li’s method (Li and Ngan, 2007). Here, both objective and subjective comparisons are provided. The test images we use are obtained from World Wide Web. There are totally 117 low DOF images along with the manually labeled ground-truth masks created by us.

Fig. 4 illustrates some representative results obtained by different methods. The images in Fig. 4(a) are named from top to bottom as Dog, Bee, Bird, Strawberry, Beckham and Flower, respectively. Fig. 4(b) gives the corresponding ground truths, where the white region corresponds to the focused object while the black region...
corresponds to the blurry background. Fig. 4(c)–(e) present the results generated in process of our method, including the focus maps, the initial segmentations by the thresholding scheme and the final segmentations by graph cut. In these images, the background is displayed in blue color. We can see that thresholding over the focus map provides the initial segmentation, though coarsely but effectively. And the graph cut optimization has accomplished the refinement elegantly, the errors in the initial segmentation are fixed and accurate results are obtained.

Fig. 4(f) and (g) show the results obtained by Kim’s method and Li’s method. As for Kim’s method, some focused objects in the test images are not correctly identified. In the dog image, a part of the dog’s body is misclassified as the background. The same error can be found in the other images as well, leading to the holes in the bird image and the missing stalk in the flower image. On the other hand, Li’s method suffers less from the missed detection, but it also fails to eliminate the false positives, such as the green region near the petal in the bee image. Compared with these two methods, our method generates more complete objects and more clear boundaries for most of the test images.

For an objective comparison, we compute the segmentation errors for the results over all the test images, using the measurement criterion which has been adopted in (Kim, 2005; Li and Ngan, 2007). It is defined as the proportion of misclassified pixels to the total number of the foreground pixels in the ground truth. Let A be the focused object in the segmentation result, and B be the one in the ground-truth mask, we can get the form of the error measurement as follows:

$$\text{error} = \frac{|A \cup B| - |A \cap B|}{|B|}.$$  \hfill (22)

where $|\cdot|$ denotes set cardinality.

Fig. 5 shows the distribution of the resulting segmentation errors for all 117 test images. It can be seen that our method has more results at the segmentation errors of 0.036 and 0.107 than Kim’s method and Li’s method. Table 1 lists the segmentation errors of the images displayed in Fig. 4. It is clear that our method is comparable to the other two methods.

In addition, we compute average precision, recall, and F-Measure over the ground-truth database, with the F-Measure defined as:

$$F_2 = \frac{(1 + \alpha) \times \text{Precision} \times \text{Recall}}{\alpha \times \text{Precision} + \text{Recall}}.$$  \hfill (23)

According to Achanta et al. (2009), we can set $\alpha = 0.3$ to weigh precision more than recall. It shows that our method achieves the highest precision rate (i.e., 91.2%) and F-Measure (i.e., 85.8%) compared with Kim’s method and Li’s method, which leads to 0.3% and 3.4% gains in precision and 1.8% and 1.2% gains in F-Measure, respectively.

To better show the performance of our method, we run the algorithm over some images which have much lower defocus degree of background versus foreground than the 117 test images used earlier. Some examples are presented in Fig. 6. The first row and the second row show two images whose focused objects have been segmented successfully. However, the third row illustrates a failure case, in which the focused football player is not completely
extracted and the other blurry player with similar colors is misclassified as the foreground. It is shown that our method may fail when the defocused background has a similar color distribution with the focused foreground, since the refinement step relies on graph cut whose performance is relative to the math model of the foreground/background. More generally, when the background is not blurred too much and has the similar defocus degree as the foreground, it is still a challenging task to segment the foreground accurately.

Note that the code of our method is implemented in Matlab, and the testing computer has a 2.60 GHz Pentium CPU with 2.00 GB of RAM, the experiment takes an average of about 7 s to segment a low DOF image of the size 400 × 300, which is comparable to Kim’s method (about 5 s on average) and much faster than Li’s method (about 40 s on average).

4. Conclusion

In this paper, we have investigated the Amplitude Decomposition Model, which indicates that the amplitude spectrum of the low DOF image can be decomposed into two components, i.e., the amplitude of its totally defocused version and the high-frequency difference amplitude of its focused regions. The model provides a foundation for focused object detection in low DOF images. Then a detection method of low complexity is proposed and utilized to assist focused object extraction. Experiments have been performed to evaluate the proposed method, the results shows that our method is comparable to the state-of-the-art methods.

Acknowledgments

This work was partially supported by NSFC(No. 60972109 and 61101091), the Program for New Century Excellent Talents in University(NCET-08-0090), and the PhD Programs Foundation of Ministry of Education of China(No. 20111085110002).

References


