Automatic segmentation of focused objects from images with low depth of field

Zhi Liu\textsuperscript{a,b,*}, Weiwei Li\textsuperscript{a}, Liquan Shen\textsuperscript{a,b}, Zhongmin Han\textsuperscript{a}, Zhaoyang Zhang\textsuperscript{a,b}

\textsuperscript{a}School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China
\textsuperscript{b}Key Laboratory of Advanced Display and System Application, Shanghai University, Ministry of Education, Shanghai 200072, China

\textbf{A R T I C L E   I N F O}

Article history:
Received 6 January 2009
Received in revised form 19 November 2009
Available online 27 November 2009

Communicated by M. Couprie

Keywords:
Focused object segmentation
Low depth of field
Focus energy estimation
Saliency map
Boundary linking
Image matting

\textbf{A B S T R A C T}

In this paper, we propose an automatic segmentation approach to extract focused objects from images with low depth of field (DOF). A focus energy map is first estimated based on the difference of high-frequency components between focused region and defocused background, and is exploited to construct region/boundary saliency maps on the basis of a pre-segmentation result by watershed transform. Then region/boundary masks for focused object are generated by entropy thresholding and flood filling, and an efficient boundary linking method is proposed to obtain closed region/boundary masks, which are exploited to reasonably generate a trimap containing seed regions for focused object and defocused background, and uncertain regions, respectively. Finally, the trimap is used as the input to an image matting model, which is utilized to classify the pixels in the uncertain regions to obtain an accurate focused object segmentation result based on the estimated alpha matte. Experimental results for a variety of low DOF images demonstrate the good segmentation performance of the proposed approach.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Images featured with low depth of field (DOF) constitute a major image category, and widely exist in professionally captured images, and videos from films and commercials. Low DOF is a photographic technique commonly used by cameraman to make the object of interest in sharp focus, while the background is blurred being out-of-focus (Kelby, 2007). From the viewpoint of optics, the range of distance in front of and behind the focused object is called the DOF. The images with low DOF clearly reflect the intention of cameraman, and also help observers to understand the depth information in a 2D image. Definitely, the semantic objects in low DOF images are the objects of interest by the cameraman, i.e., the focused objects, and therefore semantic object segmentation from such a specific class of images, i.e., low DOF images, is a well-defined problem, in which the focus information implies the most important cue for segmenting semantic objects. The main feature of low DOF images is that the focused regions are abundant in high-frequency information due to detailed edges and textures, while the defocused background regions contain less high-frequency information due to the blurring effect. Therefore, it is possible to distinguish the focused objects from the defocused background using the discrepancy on high-frequency information.

There have been a number of methods proposed for segmenting focused objects from low DOF images. The boundary based method in (Tsai and Wang, 1998) extracts the object boundaries by measuring the amount of defocus degree at each edge pixel. This method can accurately segment man-made objects with clear boundaries, but cannot deal with the disconnected boundaries of natural objects by its edge linking scheme. A series of region based segmentation methods have been presented in (Won et al., 2002; Kim, 2005; Wang et al., 2001; Ye and Lu, 2002; Li and Ngan, 2007), which are generally based on the detection of regions with high-frequency content. Different statistical features, such as local variance (Won et al., 2002) and forth-order moments (Kim, 2005; Wang et al., 2001; Ye and Lu, 2002) of high-frequency coefficients in the wavelet domain, are exploited to measure the degree of focus for each pixel. The method in (Kim et al., 2007) is further extended to block based processing for speedup. However, the above region based methods usually require a sufficiently blurred background for a reliable segmentation of focused objects. Due to the energy compaction property of wavelet transform, the use of several wavelet coefficients (Wang et al., 2001; Ye and Lu, 2002) is more prone to cause the detection of an incomplete focused object. In a recent work (Li and Ngan, 2007), a saliency map indicating the focus degree is generated based on a reblurring model, then bilateral and morphological filters are used to smooth salient regions, and finally adaptive error control matting is utilized to refine the boundaries of focused objects. This method is suitable for segmenting focused objects from the defocused background that are

\textsuperscript{*}Corresponding author. Address: School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China. Tel.: +86 21 56331419; fax: +86 21 56331194.
E-mail address: liuzhi@staff.shu.edu.cn (Z. Liu).
not sufficiently blurred, i.e., some out-of-focus objects are still recognizable in the background. However, the global use of morphological filtering in (Li and Ngan, 2007) is prone to cause incomplete object regions by global erosion and unwanted background regions by global dilation. In this paper, we consider the advantages of both boundary and region based methods, and propose an efficient segmentation approach to extract focused objects from low DOF images. For the estimation of focus degree, we introduce the orientation angle of the gradient map to model the focus energy, and evaluate the boundary saliency and region saliency, respectively. Similarly with (Li and Ngan, 2007), we also exploit image matting to obtain focused objects with smooth and precise boundaries. However, we propose a boundary linking method by jointly considering the effect of region and boundary saliency in order to generate a more suitable trimap for image matting, and the efficiency of our approach is demonstrated by relatively higher segmentation accuracy.

The rest of this paper is organized as follows. Section 2 details the proposed focused object segmentation approach. Experimental results are presented in Section 3, and conclusions are given in Section 4.

2. Focused object segmentation approach

The proposed focused object segmentation approach consists of three stages. First, two saliency maps on the basis of region and boundary are generated based on the estimated focus energy map. Next, region and boundary masks with closed contours are obtained by a boundary linking method. Finally, image matting is performed on the generated trimap to obtain accurate focused objects. The following three subsections will detail the three stages, respectively, and the last subsection will discuss the robustness of the proposed approach.

2.1. Region/boundary saliency map generation based on focus energy map

In low DOF images, high-frequency energy is more prominent in focused regions than defocused regions because the detailed structures and textures of interested objects are clearly shown in focused regions while a blurry background is usually shown in defocused regions. As the first step of our approach, we roughly delineate focused regions by constructing the focus energy map (FEM), which actually shows the distribution of high-frequency energy in low DOF images. Let $I_b$ denote the original low DOF image converted into the $L*a*b^*$ color space, and $I_d$ denotes its luminance component. It should be noted that the following processing in this stage is performed on $I_b$, which is hypothesized by the blend of a focused image $I_f$ and a defocused image $I_d$ as follows:

$$ I_b(x,y) = I_f(x,y)I_f(x,y) + (1 - I_f(x,y))I_d(x,y) $$

where $l$ is a binary mask to distinguish the focused region $R_f$ from the defocused region $R_d$, and is defined as

$$ l(x,y) = \begin{cases} 1, & (x,y) \in R_f \\ 0, & (x,y) \in R_d \end{cases} $$

Then the original image is fed into an ideal low-pass filter to eliminate the high-frequency components. The low-pass filter can take the form of 2D Gaussian function

$$ g(x,y) = \frac{1}{2\pi\sigma^2} \exp \left( -\frac{x^2 + y^2}{2\sigma^2} \right) $$

where $\sigma$ is used to control the degree of blur to generate a blurred image $I_b$. The detailed structures and textures in the focused region $R_f$ are blurred and thus exhibit obvious difference compared with the original image, while there is a little change in the defocused region $R_d$. The blurred image $I_b$ is written as

$$ I_b(x,y) = I_f(x,y)I_f(x,y) + (1 - I_f(x,y))I_d(x,y) + g(x,y) $$

In the ideal case that no high-frequency components exist in the defocused image $I_d$, we can make the following approximation

$$ I_b(x,y) + g(x,y) \approx I_f(x,y) $$

and thus $I_b$ can be approximated by

$$ I_b(x,y) = I_f(x,y)I_f(x,y) + (1 - I_f(x,y))I_d(x,y) + g(x,y) $$

It can be seen from Eqs. (1) and (6) that for the focused region $R_f$, there is a difference between $I_b$ and $I_d$. In this paper, such a difference representing the focus energy is evaluated using the histogram of oriented gradient (HOG), which was originally introduced in (Dalal and Triggs, 2005) for human detection and can effectively emphasize the salient structures. The focus energy function $E_o$ for each pixel $(x,y)$ is defined as

$$ E_o(x,y) = \frac{\left|G(x,y)\right|}{\max(\text{HOG}(x,y))} $$

where $G$ is the gradient map calculated using a Sobel operator for the original image $I_o$. Based on the orientation angle information of the gradient map, we calculate for each pixel $(x,y)$ the corresponding histogram of oriented gradient $\text{HOG}(x,y)$, which is computed over a $11 \times 11$ window centered around $(x,y)$ using 20 bins equally partitioning the whole range $[-\pi, \pi]$ for orientation angle. Correspondingly, the parameter $\sigma$ used to control the blur degree to generate the blurred image $I_b$ is set to 5, which is a half width of the above $11 \times 11$ window, in order to control the blur effect mainly in the calculation window for each pixel $(x,y)$. The selection of window size and bin partition follows the influence analysis in (Dalal and Triggs, 2005) and its suitability is verified by our experiments.

For a pixel $(x,y)$ in a region with clear edges or detailed textures, its gradient magnitude $|G(x,y)|$ is obviously higher than that of a pixel in blurry region or homogenous region. Besides, the histogram $\text{HOG}(x,y)$ of the former pixel also shows a more uniform distribution, and thus the peak of the histogram, i.e., the denominator of Eq. (7), is lower than that of the latter pixel. The advantage of using Eq. (7) is that both the magnitude information and the orientation angle information of the gradient map are exploited to distinguish the former class of pixels from the latter class of pixels. A pixel in the focused region of $I_b$ usually belongs to the former class, and thus its focus energy evaluated using Eq. (7) is higher, while the focus energy of its co-located pixel in $I_d$ is obviously reduced due to the blurred surrounding region. Therefore, the difference between the two focus energy values is usually significant for a pixel in the focused region, while such a difference is insignificant for a pixel in the defocused region. To discriminate the former class of pixels from the latter class of pixels, the focus energy map (FEM) is defined as

$$ \text{FEM}(x,y) = |E_o(x,y) - E_o(x,y)| $$

For the original low DOF image shown in Fig. 1a, its focus energy map shown in Fig. 1b highlights some focused object regions, and darkens the defocused background region. However, only clear edges and textured regions in the focused object are highlighted in FEM, while the inner homogenous regions of the object are still darkened. Such a result shown in Fig. 1b is consistent with the above analysis, and it should be noted that generally FEM can only be exploited to roughly locate the focused object.

In order to facilitate the segmentation of focused objects, we then construct the region saliency map (RSM) and the boundary saliency map (BSM) by combining the FEM and a pre-segmentation result at region level. We use watershed transform (Vincent and Soille, 1991) to pre-segment the original image into a set of
homogenous regions, which can preserve accurate boundaries between different objects on an over-segmentation result. Specifically, the gradient map \( G \) is first morphologically dilated with a 3 × 3 cross shaped structuring element \( E \), and the dilated image is elevated by a height \( h \) to get the marker image, \( G_m = (G \ominus E) + h \).

The reconstruction of \( G \) from \( G_m \) by geodesic erosion (Vincent, 1993) is performed to obtain a simplified gradient map, \( G_s(G,E) \), in which insignificant local minima are removed by the reconstruction operation \( G_s(G,E) \). The watershed transform is finally applied on \( G_s \) to obtain the pre-segmentation result. The parameter \( E \) is selected as the smallest symmetrical shaped structuring element in order to introduce the least discrepancy on object boundaries into the gradient map \( G_s \). The other parameter \( h \) is set to 2 by our experiments, and this smaller value leads to a reasonable pre-segmentation result with over-segmentation, which suffices to preserve focused object boundaries. The pre-segmentation result of the image in Fig. 1a is shown in Fig. 1c, in which each black line represents the boundary between two adjacent regions. It should be noted that a collection of algorithms such as mean shift and graph based segmentation can also be used to generate an over-segmentation result, while we select watershed transform due to its lower computation complexity.

The RSM is generated on the basis of segmented region, specifically, the region saliency for each segmented region \( R_i \) is defined as

\[
RS(R_i) = \frac{\sum_{(x,y) \in R_i} FEM(x,y)}{|R_i|},
\]

where \( |R_i| \) represents the area of the region \( R_i \).

The BSM is generated on the basis of boundary line, specifically, for each boundary \( B_{ij} \) separating the two adjacent regions \( R_i \) and \( R_j \), its boundary saliency is defined as

\[
BS(B_{ij}) = \frac{\sum_{(x,y) \in B_{ij}} FEM(x,y)}{|B_{ij}|},
\]

where \( |B_{ij}| \) represents the length of the boundary \( B_{ij} \).

Then RSM and BSM at the pixel level are generated as follows:

\[
RSM(x,y) = RS(R_i), \quad \forall (x,y) \in R_i
\]

\[
BSM(x,y) = BS(B_{ij}), \quad \forall (x,y) \in B_{ij}
\]

Based on the pre-segmentation result in Fig. 1c and the focus energy map in Fig. 1b, the corresponding RSM and BSM are shown in Fig. 1d and e, respectively. It can be seen from Fig. 1d that nearly all defocused background regions are dimmed as expected, but some regions of the focused object are also dimmed in RSM. Similarly in Fig. 1e, most boundaries that separate the focused object and the background are highlighted as expected, but some of them are not sufficiently highlighted. In the next subsection, we will exploit RSM and BSM to obtain suitable masks for focused objects on the basis of region and boundary, respectively.

### 2.2. Region/boundary mask for focused object

The initial detection of regions and boundaries for focused objects is performed by thresholding RSM and BSM, respectively. The threshold is automatically determined using the entropic thresholding technique (Cheng et al., 2000). The FEM is first scaled into the range of \([0,255]\), and a 256-bin histogram is generated to record the number of pixels \( n_k \) for each gray level \( k \). Given a threshold \( T_e \), the probability for focused object pixels is defined as

\[
P_f(k) = \frac{n_k}{\sum_{k=1}^{255} n_k}, \quad T_e + 1 \leq k \leq 255
\]

and the probability for defocused background pixels is defined as

\[
P_d(k) = \frac{n_k}{\sum_{k=T_e}^{255} n_k}, \quad 0 \leq k \leq T_e
\]

Then the entropies for the above two pixel classes using the threshold \( T_e \) are defined as

\[
ET_f(T_e) = -\sum_{k=T_e+1}^{255} P_f(k) \log P_f(k)
\]

\[
ET_d(T_e) = -\sum_{k=0}^{T_e} P_d(k) \log P_d(k)
\]

The optimal threshold \( T_e \) is determined using the following criterion:

\[
T_e = \arg \max_{T_e=0.1 \ldots 255} \{ ET_f(T_e) + ET_d(T_e) \}
\]

For the RSM and BSM shown in Fig. 1d and e, the initial region mask (IRM) and the initial boundary mask (IBM) for the focused object are obtained using \( T_e \). For the IMR is shown in Fig. 2a and b, respectively. Based on the IBM in Fig. 2b, the connected component analysis is exploited to identify the largest connected component, which contains a majority of connected boundaries of the focused object. The connected components far from the largest connected component are highly likely to locate in the defocused background, and thus can be removed as irrelevant noisy boundaries. Let \( C_{C_0} \) denotes the largest connected component, any connected component \( C_i \) is removed from IBM when the following criterion is satisfied

\[
\min_{p \in C_i, q \in C_{C_0}} \text{dist}(p,q) \leq T_d
\]
where \( \text{dist}(p, q) \) is the Euclidean distance between the pixel \( p \) and
the pixel \( q \). The threshold \( T_d \) is set to 5, which is sufficient to reliably
remove those noisy boundaries in the defocused background.

The flood filling method is then performed on IBM to fill those
segmented regions that are completely closed by the detected
boundaries. The filled boundary mask (FBM) for Fig. 2b is shown
in Fig. 2c, in which closed regions are shown with white pixels. It
can be seen from Fig. 2c that a majority part of the focused object
is identified in FBM, but some object regions are inevitably missed
in FBM due to some unclosed boundaries. The IRM shown in Fig. 2a
is also refined to obtain the filled region mask (FRM) shown in
Fig. 2d by filling inner boundaries between regions and eliminating
corresponding irrelevant noisy regions identified in IBM.

It can be seen from Fig. 2c and d that some unclosed boundaries
usually exist in FBM and missed object regions also occur in FRM.
In order to obtain a more accurate representation for focused ob-
jects, a boundary linking method is proposed to generate a closed
boundary mask (CBM). The following will detail the proposed
method to generate closed regions for those unclosed boundaries
in FBM. The closed regions generated by boundary linking are con-
sidered as uncertain regions, which could possibly belong to the fo-
cused object or defocused background. Each unclosed boundary in
FBM is checked with the following three situations (self-closing,
multip-closing and hyper-closing) in turn to form a closed region.
The schematic picture in Fig. 3 is used to illustrate the three situa-
tions as follows.

(a) **Self-closing.** For the unclosed boundary \( L_1 \), there are three
other boundaries connected with \( L_1 \) at the junction point
\( p_1 \). The boundary with the highest boundary saliency value
is selected as the candidate linking boundary (marked as the red line). With the addition of the candidate linking
boundary, \( L_1 \) becomes a part of the closed region \( R_1 \), and
thus \( R_1 \) is selected as an uncertain region.

(b) **Mutual-closing.** If the unclosed boundary cannot conform to
the situation of *self-closing*, it is then tested with the situa-
tion of *mutual-closing* described as follows. For each
unclosed boundary, its candidate linking boundary with
the highest boundary saliency value is first selected. Exclud-
ing the boundaries conforming to the situation of *self-closing*,
we try to find any closed region that can be composed by
multiple unclosed boundaries and their candidate linking
boundaries. For the example of the two unclosed boundaries
\( L_2 \) and \( L_3 \), the corresponding two candidate linking bound-
aries are marked as the two red lines connecting at \( p_2 \) and
\( p_3 \), respectively. With the addition of the two candidate link-
ing boundaries connecting at the junction point \( p_4 \), \( L_2 \) and
\( L_3 \) become a part of the closed region \( R_2 \), and thus \( R_2 \) is
selected as an uncertain region accordingly.

(c) **Hyper-closing.** If the unclosed boundary cannot conform to
the above two situations, it is then tested with the situation of *hyper-closing* described as follows. From the two adjacent
regions that are separated by the unclosed boundary, the
region with the higher region saliency value is selected as the
closed region. For example, the unclosed boundary \( L_4 \)
cannot conform to the above two situations, and the corre-
sponding two adjacent regions separated by \( L_4 \) are \( R_3 \) and
\( R_4 \), respectively. If the region saliency value of \( R_3 \) is higher
than that of \( R_4 \), the closed region \( R_3 \) is selected as the uncer-
tain region.

The above boundary linking method is performed on the FBM
shown in Fig. 2c, and the corresponding CBM is shown in Fig. 2e,
in which the unclosed boundaries shown in Fig. 2b are completely
linked. The closed regions obtained by boundary linking are
marked as gray regions in Fig. 2f, and they will be considered as
uncertain regions for further processing in the next subsection.

### 2.3. Focused object extraction by image matting

In order to obtain an accurate segmentation of focused objects,
the above obtained region and boundary masks are exploited to
generate a trimap, which will be used as the input to an image
matting model (Ruzon and Tomasi, 2000; Levin et al., 2008). The pixels in the trimap are classified into three regions, i.e., TRIF, TRIU, and TRID, representing seed regions for focused object, seed regions for defocused background, and uncertain regions, respectively, are defined as follows:

\[
TRIF = FRM
\]

\[
TRID = \Phi(FCBM)
\]

\[
TRIU = \overline{(TRIF \cup TRIU)}
\]

where FCBM is the region mask obtained by flooding filling on the original area of CBM. Since the matting is mainly exploited to refine the object boundaries, a smaller dilation of 5% area is enough for this purpose. The uncertain region TRIU is the remainder of the whole image excluding TRIF and TRID. The generated trimap for the example in Fig. 2 is shown in Fig. 4a where white, black and gray regions denote TRIF, TRIU, and TRID, respectively.

Under the framework of image matting, the original low DOF image \(I\) is assumed to be a composite of a foreground image \(F\) and a background image \(B\). The color of each pixel is assumed to be a linear combination of the corresponding foreground color and background color, and is defined as

\[
I(x, y) = \alpha(x, y)F(x, y) + (1 - \alpha(x, y))B(x, y)
\]

where \(\alpha(x, y) \in [0, 1]\) is the alpha matte value indicating the foreground opacity for the pixel \((x, y)\). Based on the above generated trimap, the alpha matte value for each pixel in TRIF and TRIU is initialized with 1 and 0, respectively. The alpha matte values for pixels in TRID are solved by the method in (Levin et al., 2008), which provides a globally optimal alpha matte by solving a sparse linear system of equations, rather than performing iterative nonlinear estimation by alternating foreground and background color estimation with alpha estimation.

Based on the trimap in Fig. 4a, the estimated alpha matte is represented by the gray-level image shown in Fig. 4b. The alpha matte is then thresholded to obtain a binary segmentation of foreground (focused object) and background (defocused background). In our implementation, the threshold for the alpha matte is set to 0.5. It is observed from our experiments that the final segmentation results are not very sensitive to the selected threshold, since the obtained alpha matte is usually distinguishable for a binary classification. After thresholding the alpha matte in Fig. 4b, the finally extracted focused object with accurate and smooth boundaries is shown in Fig. 4c.

2.4. Discussion

The three stages of the proposed focused object segmentation approach are closely interrelated, and the results of the former stage influence the results of the latter stage to some extent. Specifically, the focus energy map and the pre-segmentation result in the first stage are the two factors that affect the intermediate results and the final result generated in the latter two stages. With a suitable focus energy map and a reasonable pre-segmentation result, the performance of the latter two stages is robust to extract the focused objects with acceptable quality. The former factor, focus energy map, is directly related with the amount of defocus exhibited in the original low DOF images, i.e., the defocus degree of background versus object has a direct effect on the quality of finally extracted objects, and we will give a statistical analysis in the section of experimental results.

In this subsection, the effect of the latter factor, pre-segmentation result, on different stages of the proposed approach is analyzed using the examples shown in Figs. 5 and 6. Starting with different pre-segmentation results, the intermediate results generated in the three stages will show a discrepancy and affect the finally extracted objects to some extent. As we have mentioned above, a suitable pre-segmentation result for the proposed approach is with somewhat over-segmentation to preserve the object boundaries, and this can be fulfilled by a collection of image segmentation methods. Compared with the pre-segmentation result
in Fig. 1c, an excessive over-segmentation result and an under-segmentation result are shown in Figs. 5a and 6a, respectively. The pre-segmentation result shown in Fig. 5a is generated by the watershed transform, in which the parameter $h$ is set to 1 (a smaller value), and region merging is performed on Fig. 5a to generate the under-segmentation result shown in Fig. 6a. It can be seen that all boundaries of the focused object are also well preserved in Fig. 5a, while not all boundaries of the focused object are preserved in Fig. 6a, for example, the boundary at the bottom-right part of the object is missed in Fig. 6a. Using the two pre-segmentation results and the same focus energy map in Fig. 1b, the correspondingly generated RSMs and BSMs are shown in Figs. 5b and c, and 6b and c, respectively, and it can be seen that a majority of regions and boundaries belonging to the focused object are still reasonably highlighted in both results. It should be noted that some regions and boundaries of the focused object generally cannot be sufficiently highlighted in RSM and BSM due to a lower defocus degree and/or a lower contrast around some object boundaries, and the latter two stages are thus exploited to efficiently process such RSM and BSM to extract the focused object with a reasonable quality.

Based on the above RSMs and BSMs, the two important intermediate results, i.e., IBMs and CBMs generated in the second stage are shown in Figs. 5d and e, and 6d and e, respectively. We can see that some object boundaries are missed in both IBMs, but the proposed boundary linking method can generate reasonably closed boundary masks. It is obvious that the CBM in Fig. 5e delineates nearly all boundaries of the focused object, while the CBM in
Fig. 6e is also a reasonable result generated from such an under-segmentation result in Fig. 6a.

Based on the two CBMs in Figs. 5e and 6e, the correspondingly generated two trimaps in the third stage are shown in Figs. 5f and 6f, respectively. The estimated alpha mattes and the finally extracted focused objects are shown in Figs. 5g and h, and 6g and h, respectively. It can be seen that Fig. 5g and h are nearly the same as Fig. 4b and c, i.e., the focused object is also extracted with good visual quality starting from a different pre-segmentation result with over-segmentation. However, for the pre-segmentation result with under-segmentation, the extracted object in Fig. 6h contains some redundant background regions.

As shown in Figs. 5 and 6, we can firstly conclude that a suitable pre-segmentation result for our approach should preserve object boundaries, but the quality of the finally extracted focused object is not sensitive to different over-segmentation results. Secondly, the proposed boundary linking method is adequate to fulfill the task of efficiently closing disconnected boundaries, while the quality of the generated CBM actually depends on the two factors mentioned above. Finally, based on the intermediate results of previous two stages, the proposed trimap generation method is adequate to provide a reasonable trimap for the matting method to extract the focused object with a reasonable quality.

3. Experimental results

The proposed focused object segmentation approach is evaluated on a total of 80 low DOF images, which are selected from the Corel image database and downloaded from the Internet. The focused object segmentation results for several representative images are shown in Figs. 7 and 8, respectively. In both figures, the original images, the focused objects manually segmented as the ground truth, and the focused objects segmented by the proposed approach are shown from the 1st to the 3rd column in turn, and two measures, the defocus degree of background versus object and the segmentation error rate, are shown in the last two columns.

In order to objectively evaluate the segmentation performance, we adopt the pixel-based quality measure (Wollborn and Mech, 1998) to define the segmentation error rate as follows:

\[
e(\mathbf{M}_\text{seg}, \mathbf{M}_\text{gt}) = \frac{\sum_{(x,y)} \mathbf{M}_\text{seg}(x,y) \odot \mathbf{M}_\text{gt}(x,y)}{\sum_{(x,y)} \mathbf{M}_\text{gt}(x,y)}
\]

where \(\mathbf{M}_\text{seg}\) and \(\mathbf{M}_\text{gt}\) are the binary mask for the segmented focused object and the manually segmented ground truth, respectively, and \(\odot\) is the binary "XOR" operation.
From the view of the focused object segmentation, we define the defocus degree of background versus object for the original low DOF image as follows:

\[
D = \frac{\sum_{(x,y)} \text{FEM}(x,y) \cdot M_{gt}(x,y)}{\sum_{(x,y)} \text{FEM}(x,y) \cdot [1 - M_{gt}(x,y)] / \sum_{(x,y)} [1 - M_{gt}(x,y)]}
\]

where the numerator is the average focus energy value for the focused object region in the ground truth, and the denominator is the average focus energy value for the background region in the ground truth. A higher value of the above defined defocus degree measure usually indicates a sharply focused object with a sufficiently defocused background.

It can be seen from the former four examples in Fig. 7 that the sharply focused objects with a sufficiently blurred background or a uniform background can be accurately extracted. In the latter two examples in Fig. 7, although the defocused background shows a similar color with some part of the focused object, the defocus degree values for the two images are moderate, and the proposed approach can extract the focused objects with relatively lower segmentation error rates.

Compared with the examples with relatively higher value of defocus degree (usually greater than 5) in Fig. 7, some examples with relatively lower value of defocus degree are shown in Fig. 8. In the former two examples, there are still some recognizable objects in the defocused backgrounds, but these defocused objects do not appear as false alarms in the segmentation results due to considerable distances to the focused objects in the depth direction. For the latter three examples in Fig. 8, the user-interested objects, i.e., big pink flowers, white crane and black bear, are definitely included in the ground truth, and those regions that locate in the depth range of the user-interested objects are then added into the ground truth. The above criterion for defining the ground truth is somewhat strict, but it avoids ambiguity on ground truth used for objective evaluation of object segmentation performance. In the 3rd example, some regions around the flowers are not totally out-of-focus and thus appear in our segmentation result. In the 4th and the 5th example, the very low values of defocus degree (less than 2) indicate that the defined focused object regions in the ground truths are not distinct from background regions, and some background regions surrounding the user-interested objects also show a similar color with the adjacent focused...
object regions. The above factors lead to relatively higher segmentation error rates for the last two examples.

We can clearly see from Figs. 7 and 8 that the segmentation error rate is related with the defocus degree, which also indicates the difficulty level to extract a perfect user-interested focused object to some extent. Using Eqs. (22) and (23), we calculate the segmentation error rate and the defocus degree for all test images, and make a statistic on the average segmentation error rate for each class of images within a specified range of defocus degree. The whole range of defocus degree is divided into 6 specified ranges. Specifically, the range of [1, 11] is divided into 5 specified ranges with an equal interval of 2, and the range of [11, +∞) is used as the last one. Fig. 9 plots the average segmentation error rates for different ranges of defocus degree. It can be seen from Fig. 9 that the average segmentation error rate decreases with the increase of defocus degree. Furthermore, we can notice that there is an obvious reduction of the average segmentation error rate when the defocus degree is greater than 3, and it is rather stable around 0.1 within the defocus degree range of [5, 11]. Based on the observations from Fig. 9, we can conclude that the proposed approach is robust to segment the focused objects with lower segmentation error rates from a variety of low DOF images, whose defocus degree values are usually greater than 3. We should also note that for some low DOF images with very low defocus degree values (usually less than 3), e.g., the last two examples in Fig. 8, it is hard to accurately extract pure user-interested focused objects without including some surrounding background regions, which locate within the approximate depth range of the sharply focused objects.

The proposed approach is compared with the segmentation approach recently proposed in (Li and Ngan, 2007). The segmentation results for comparison are shown in Fig. 10, in which the original low DOF images are shown in the 1st column, the focused objects manually segmented as the ground truth are shown in the 2nd column, and the focused objects extracted by Li and Ngan’s approach and the proposed approach are shown in the 3rd and the 4th column, respectively. For a more clear inspection on darker regions of some focused objects, the background regions are set to white in Fig. 10. For low DOF images with one sharply focused object (see the 1st and the 2nd rows), both approaches achieve satisfactory segmentation results. For the image in the 3rd row without sufficiently defocused background, the focused object extracted by Li and Ngan’s approach also contains a portion of background regions, while the extracted focused object by our approach do not introduce such false alarms. For the image in the 4th row, a part of playfield is still in focus, and thus some playfield regions are unavoidably contained in both results. However, the two players extracted by our approach are more complete and show a better visual quality. For the last image, the regions of scattered flowers and the defocused background show a similar color with the focused objects, and thus these regions appear in the results obtained by both approaches. The data of defocus degree for the images in Fig. 10, and the segmentation error rates achieved by both approaches are shown in Table 1. We can see that our approach achieves a lower segmentation error rate for all five images with a wide range of defocus degree. Both subjective observation and objective evaluation demonstrate the better segmentation performance of the proposed approach.

The proposed approach is further compared with another segmentation approach proposed in (Kim, 2005), which is suitable

Table 1

<table>
<thead>
<tr>
<th>Image number</th>
<th>Defocus degree</th>
<th>Li and Ngan</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.65</td>
<td>0.074</td>
<td>0.050</td>
</tr>
<tr>
<td>2</td>
<td>6.80</td>
<td>0.048</td>
<td>0.039</td>
</tr>
<tr>
<td>3</td>
<td>7.51</td>
<td>0.241</td>
<td>0.114</td>
</tr>
<tr>
<td>4</td>
<td>8.46</td>
<td>0.125</td>
<td>0.093</td>
</tr>
<tr>
<td>5</td>
<td>3.16</td>
<td>0.303</td>
<td>0.289</td>
</tr>
</tbody>
</table>
for segmenting focused objects from gray-level low DOF images. As shown in Fig. 11, the gray-level low DOF images are shown in the 1st column, the manually segmented ground truths are shown in the 2nd column, and the focused object segmentation results using Kim’s approach and our approach are shown in the 3rd and the 4th column, respectively. Table 2 lists the data of defocus degree for each image in Fig. 11, and the segmentation error rates achieved by both approaches. It can be seen from Table 2 that the defocus degree for these images are moderate, and our approach achieves a better segmentation quality than Kim’s approach for all images except for the 1st image in Fig. 11, in which the antennae of the butterfly are not preserved in our result. We can observe from the 2nd row of Fig. 11 that our segmentation result correctly contains some grass regions in focus, and in the last row of Fig. 11, the defocused background region among the football players is correctly eliminated in our result. The subjective observations on Fig. 11 accord well with the segmentation error rates listed in Table 2.

4. Conclusion

In this paper, we have presented an efficient focused object segmentation approach for low DOF images. First, the difference of focus energy between focused regions and defocused regions is highlighted using the histogram of oriented gradient, and region/boundary saliency maps are constructed by considering the advantages of both boundary and region based methods. Second, a boundary linking method that jointly considers the effect of region and boundary saliency is proposed for generating a trimap suitable for image matting. Finally, based on the estimated alpha matte, the focused objects are segmented with smooth and precise boundaries. Experimental results demonstrate that our approach can effectively segment focused objects from low DOF images with better segmentation performance.

Acknowledgements

The authors are grateful to the anonymous reviewers and the associate editor for their valuable comments, which have greatly helped us to improve this work. The authors also thank Dr. Hongliang Li and Prof. King Ngi Ngan for providing their help to make a comparison. This work is supported in part by National Natural Science Foundation of China under Grant No. 60602012, in part by Shanghai Educational Development Foundation under Grant No. 2007CG53, and in part by Innovation Program of Shanghai Municipal Education Commission.(No. 09YZ02).

References


Table 2
Objective evaluation for comparison with Kim’s approach.

<table>
<thead>
<tr>
<th>Image number</th>
<th>Defocus degree</th>
<th>Kim error rate</th>
<th>Proposed error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.53</td>
<td>0.079</td>
<td>0.085</td>
</tr>
<tr>
<td>2</td>
<td>6.46</td>
<td>0.187</td>
<td>0.102</td>
</tr>
<tr>
<td>3</td>
<td>9.24</td>
<td>0.133</td>
<td>0.111</td>
</tr>
<tr>
<td>4</td>
<td>7.53</td>
<td>0.219</td>
<td>0.177</td>
</tr>
</tbody>
</table>