Multiresolution Object-of-Interest Detection for Images with Low Depth of Field*

Jia Li† James Ze Wang‡ Robert M. Gray§ Gio Wiederhold¶
Stanford University, Stanford, CA 94305, USA

Abstract

This paper describes a novel multiresolution image segmentation algorithm for separating sharply focused objects-of-interest from other foreground or background objects in low depth of field (DOF) images, such as sports, telephoto, macro, and microscopic images. The algorithm takes a multiscale context-dependent approach to segment images based on features extracted from wavelet coefficients in high frequency bands. The algorithm is fully automatic in that all parameters are image independent. Experiments with the algorithm on more than 100 low DOF images have shown results close to the human segmentation of these images. Besides high accuracy, the algorithm also provides high speed. A 768×512 *pixel image can be segmented within two seconds on a Pentium Pro 300MHz PC.*

1 Introduction

Unsupervised image segmentation [6] is invariably one of the most challenging problems in the field of computer vision. It is also crucially important in applications such as target recognition [2], image understanding, and content-based image database indexing and retrieval [5, 7, 15]. In this paper, we focus on the segmentation of low depth of field (DOF) images. Low DOF is an important technique widely used by professional photographers for various types of images, such as telephoto images, to emphasize a certain object. It is also a key technique for microbiologists to understand the 3-D structure within a specimen under a high-power microscope.

Normal human vision is nearly infallible in segmenting sharply focused objects-of-interest in a low DOF image. Most currently available segmentation algorithms, however, fall far short of human performance in this task. To mimic the human perception which uses both global and local information in segmentation, a novel multiresolution segmentation algorithm is developed for low DOF images. The algorithm aims at separating sharply focused objects-of-interest from other foreground or background objects. It is fully automatic in that all parameters are image independent.

2 Low Depth of Field

In this section, we briefly describe the concept of depth of field in photography. For convenience, we call the sharply focused object-of-interest the OOI, and the other out-of-focused foreground and background objects the background. In actuality, some background objects may be closer to the camera than the OOI .

2.1 Depth of Field

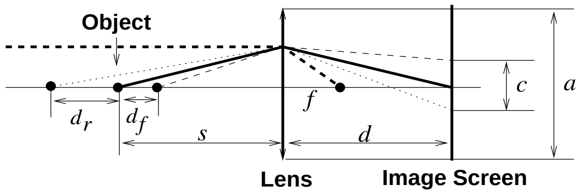


Figure 1. The optical construction of a typical camera.

Depth of field (DOF) is the range of distance from a camera that is acceptably sharp in the photograph [1, 13]. A typical camera is an optical system

*This work was supported in part by the National Science Foundation under NSF Grant No. MIP-9706284. We would like to thank Oscar Firschein, Visiting Scholar at Stanford University, and various researchers including Martin A. Fischler and Quang-Tuan Luong of the SRI AI Center for the valuable discussions in computer vision and photography.

†Department of Electrical Engineering.
Email: `jiali@isl.stanford.edu`
‡Department of Computer Science and Department of Medical Informatics.
§Department of Electrical Engineering.
¶Department of Computer Science.

containing a lens and an image screen. The lens creates images in the plane of the image screen, which is normally parallel to the lens plane. Figure 1 illustrates the optical construction of a typical camera.

Denote the focal length of the lens by f and its diameter by a. Denote the aperture f-stop number for this photo by p. Then $f = ap$. Suppose the image screen is at distance d from the lens and the object is at distance s from the lens. If the object is in focus, then the Gaussian thin lens law holds:

$$\frac{1}{s} + \frac{1}{d} = \frac{1}{f} \ .$$

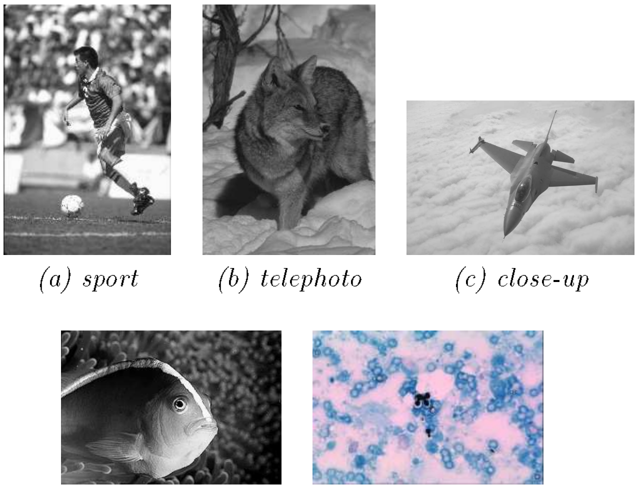
A point closer or farther away from the lens than s is imaged as a circle, as shown in Figure 1. Assume the largest circle that a human can tolerate, namely, the circle of minimum confusion, has a diameter of c. A point is considered sharp if and only if the image of the point is smaller than the circle of minimum confusion. As shown in Figure 1, d_f and d_r are the front and the rear DOF limits, respectively.

By simple geometry, it can be shown that

$$d_f = \frac{scp(s-f)}{f^2 + cp(s-f)} \ , \ d_r = \frac{scp(s-f)}{f^2 - cp(s-f)} \ .$$

Usually the size of the circle of minimum confusion is fixed for a given image size. For a fixed circle of minimum confusion, we conclude from the above equations that larger aperture, closer object distance, or longer focal length, leads to lower DOF.

2.2 Low DOF and Photography



(a) sport *(b) telephoto* *(c) close-up*

(d) macro *(e) microscopic*

Figure 2. Types of images with low DOF.

wide angle *small aperture*

Figure 3. High DOF images.

With low DOF, only the OOI is in sharp focus, whereas background objects are typically blurred to out-of-focus. Photographers often use low DOF to create a sense of depth in two dimensional photograph. Examples are shown in Figure 2. If we look at a typical sports image such as Figure 2(a), the player is much sharper than the background crowds. We know immediately that the player is in the foreground and is the OOI. As shown in Figure 2 (b)$\sim$(d), low DOF is also used in telephoto, closeup, and macro photography to distinguish the object-of-interest and the background. Due to the special focal length of microscopic lenses, images obtained from microscopes are usually of low DOF (Figure 2(e)). For comparison, Figure 3 shows that in high DOF images, both the OOI and the background are sharply focused.

2.3 Low DOF and Image Segmentation

The normal human vision system (HVS) is nearly infallible in understanding both low DOF and high DOF photographs. For the case of understanding high DOF photographs, human knowledge plays a key role. For example, the HVS is capable of interpreting a lake in a scene as a flat surface. The HVS can also understand cartoon sketches when no detailed texture information is available.

On the other hand, for images with substantially distinct depths, such as images with low DOF, the HVS also uses depth information to assist image understanding. Low DOF is often used for images in which the background is distracting to viewers. This is the main reason that low DOF is an important technique for professional photographers. Low DOF microscopic imaging is also important for microbiologists who use a low DOF microscope to determine the 3-D structure of a specimen from 2-D slices.

Recent work has taken advantages of DOF in the field of computer vision and image understanding. Among others [3, 10, 11, 14, 12], Yim and Bovik [16] have explored the possibility of depth perception using a sequence of images taken with different image plane distances. Yim and Bovik [16] have also provided a detailed survey of these techniques.

DOF could be an important clue to a computer vision system for understanding images with a low DOF. In fact, a unique DOF is associated with any photographic image taken by a conventional camera. Our algorithm detects different depths in a low DOF image by analyzing wavelet coefficients and segments the image according to the DOF information.

3 Classification Features

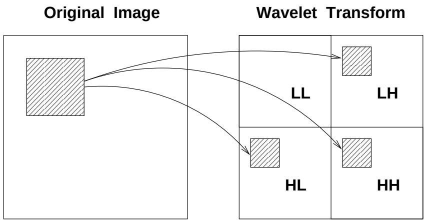


Figure 4. The wavelet coefficients at the same spatial location in the high frequency bands for a block in an image.

We divide an image into blocks and classify each block as *background* or *object-of-interest*. The classifier uses two features, the average intensity of an image block and the variance of wavelet coefficients in the high frequency bands. The average intensity is used to test how similar one block is to another. The variance of wavelet coefficients in the high frequency bands is the main feature to distinguish background and OOI. As we assume that only the OOI is in focus in a low DOF image, details in the OOI are captured but those in background are not. The details in the OOI result in larger high frequency energy in an image. We measure the high frequency energy by the variance of wavelet coefficients in the high frequency bands, i.e., the LH, HL and HH bands shown in Figure 4. For any image block, the variance of wavelet coefficients at the same spatial location in the three high frequency bands is used as a feature for the image block. We denote this feature as v. Suppose an image is specified by a set of pixels $\Im = \{(m,n),\ m = 0,...,M-1,\ n = 0,...,N-1\}$ and its wavelet coefficients are $\{w_{m,n},\ (m,n) \in \Im\}$. Without loss of generality, consider block $\{(m,n),\ m = 0,...,s-1,\ n = 0,...,s-1\}$. The wavelet coefficients for the block in LH, HL, and HH band are $\{w_{m,n}, m = 0,...,s/2-1, n = N/2,...,N/2+s/2-1\}$, $\{w_{m,n}, m = M/2,...,M/2+s/2-1, n = 0,...,s/2-1\}$, and $\{w_{m,n}, m = M/2,...,M/2+s/2-1, n = N/2,...,N/2+s/2-1\}$ respectively. The feature v is then calculated as the variance of the wavelet coefficients in all the three sets. For other shifted image blocks, their wavelet coefficient blocks for calculating v are shifted correspondingly, as shown in Figure 4.

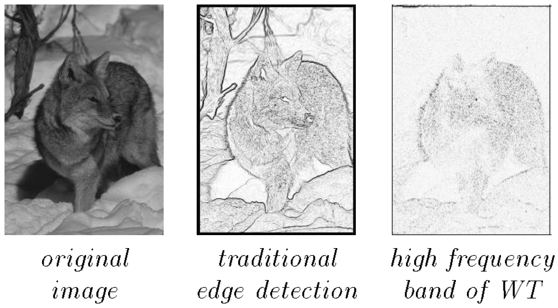


Figure 5. Comparison of the traditional edge detection and wavelet transform.

In our current implementation, the Haar wavelet transform is used. We expect Daubechies' wavelet transforms [4] with short length filters will give similar results. Figure 5 shows a comparison between traditional edge detection and a wavelet transform. The OOI stands out more distinctly in the high frequency band of the wavelet transform than that does in the traditional edge detected image.

4 The Algorithm

The classification algorithm consists of three steps:

1. initial classification at the lowest scale
2. recursive process to adjust the crude classification result using a multiscale approach
3. post-processing to obtain smooth boundaries and to remove small isolated regions

As shown in Figure 6, we start with a large block size. A crude classification is performed with the large blocks. At every increased scale, the blocks are subdivided into four *child* blocks, forming a quad-tree structure. Child blocks inherit the classes of parent blocks as their initial classes. The classifier then adjusts the classes of the child blocks according to their features and their context, which is represented by the statistics of their neighboring blocks. After the adjustment is performed, the classifier increases the scale and repeats the previous step until the maximum scale is reached. This multiscale approach is motivated by a similar context-dependent classification structure, developed and applied to document segmentation by Li and Gray [9].

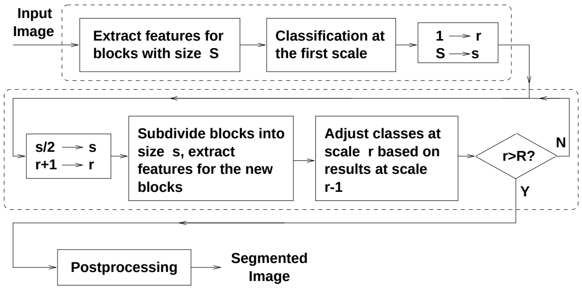


Figure 6. Flow chart of the algorithm.

4.1 Initial Classification

For the initial classification, we start with a large block size $S \times S$, which is usually set to 32×32 for an image of around 768×512 pixels in our applications. We denote the lowest scale by $r = 0$. By avoiding over-localization, large blocks provide a more distinguished feature v, defined in the previous section, for the two classes: background and OOI. Although classification performed on the large blocks is crude, we compensate this by adjusting the classes at higher scales.

Denote the blocks at scale $r = 0$ by $\mathfrak{S}^{(0)} = \{(i,j), i = 0, ..., I, j = 0, ..., J\}$ and the feature v for the blocks by $v_{i,j}, (i,j) \in \mathfrak{S}^{(0)}$. We then divide $v_{i,j}$ into two clusters using the k-means clustering algorithm [8]. One cluster represents the background, and the other represents the OOI. In the view that the OOI has a higher average v, we set the cluster with the higher average v as the OOI class. Denote the background cluster center by $v^{(0)}$, and the OOI cluster center by $v^{(1)}$. The k-means algorithm determines the class of block (i,j), $(i,j) \in \mathfrak{S}^{(0)}$, at scale $r = 0$, by $c_{i,j} = \min_{k=0,1}^{-1}(v_{i,j} - v^{(k)})^2$.

At the end of the initial classification, we delete small isolated background regions to avoid smooth regions of the OOI being mistakenly classified.

4.2 Multiscale Context Dependent Classification

The second step in the classification is to adjust the segmentation result obtained in the first step using context information through a multiscale approach. At each increased scale, the blocks are divided into a quadtree with 4 child blocks. The parent block of block (i,j) is denoted by $(\tilde{i}, \tilde{j})$, where $\tilde{i} = \lfloor i/2 \rfloor$, $\tilde{j} = \lfloor j/2 \rfloor$. Denote the set of blocks at scale r by $\mathfrak{S}^{(r)}$, $r = 1, ..., R$, where R is the maximum scale set by users. The block size at scale R is the finest resolution that a user needs. We choose R as the scale at which one block is a single pixel in our applications.

The feature v and the average intensities of blocks at scale r are evaluated. The initial classes of the blocks are set to the classes of their parent blocks. For every block which is adjacent to a block with a different class, the classifier decides whether to switch the class of one of the blocks according to their v, and the similarity to their parent blocks.

The multiscale approach is used to solve the conflict of over-localization and high resolution classification. By gradually reducing the block size, the classifier can track the more global properties of a block through its parent block. At the same time, a high resolution classification is achieved eventually. The details of the algorithm are provided in the list below.

1. Set $1 \rightarrow r$.
2. Divide blocks in $\mathfrak{S}^{(r)}$ into blocks in $\mathfrak{S}^{(r+1)}$. Set $r + 1 \rightarrow r$.
3. Calculate features $v_{i,j}$ and average intensities $x_{i,j}$ for blocks $(i,j) \in \mathfrak{S}^{(r)}$.
4. For each block $(i,j) \in \mathfrak{S}^{(r)}$ segmented as the OOI, if any of its four neighbors is segmented as background, adjust the classes as follows.
   (a) Set $0 \rightarrow k$.
   (b) For the kth neighbor block (m,n), set $flip_{background} = 1$ if one of the following conditions is satisfied
      i. The difference between the average intensity $x_{m,n}$ of block (m,n) and that of its parent block is larger than both a threshold and the difference between $x_{m,n}$ and the average intensity of the parent block of (i,j).
      ii. The feature $v_{i,j}$ is closer to the center of the OOI cluster, i.e., $(v_{i,j} - v^{(1)})^2 < (v_{i,j} - v^{(0)})^2$
   (c) For the kth neighbor block (m,n), set $flip_{OOI} = 1$ if one of the following conditions is satisfied
      i. The difference between the average intensity $x_{i,j}$ of block (i,j) and that of its parent block is larger than the difference between $x_{i,j}$ and the average intensity of the parent block of (m,n), and $|x_{i,j} - x_{\tilde{m},\tilde{n}}| < \theta$, where θ is a threshold.

ii. The difference between the average intensity $x_{i,j}$ of block (i,j) and that of its parent block is larger than the difference between $x_{i,j}$ and the average intensity of the parent block of (m,n), and $v_{i,j}$ is much closer to $v^{(0)}$.

(d) If $flip_{OOI} = 1$ and $flip_{background} = 0$, switch the class of block (i,j) to background.

(e) If $flip_{OOI} = 0$ and $flip_{background} = 1$, switch the class of block (m,n) to the OOI.

(f) If block (i,j) is switched to background, and some neighbor block is switched to the OOI, change the class of the neighbor block back to background.

5. If $r \leq R$, go to step 2; else, stop.

The algorithm above checks all the OOI blocks adjacent to background blocks and switches the classes of the block and its neighbors if needed. Since all background blocks adjacent to an OOI block are compared to the OOI block, it is thus redundant to go through all the background blocks as well and compare them with their neighboring blocks. The thresholds used in the algorithm are pre-selected and fixed for any test image. Note that the cluster centers $v^{(0)}$ and $v^{(1)}$ for background and OOI are fixed through all the scales. Also, the features $v_{i,j}$ are discarded at very high scales. When the scale is sufficiently high, the blocks shrink to very small sizes. Small blocks in both background and OOI are likely to be smooth, thus $v_{i,j}$, which are the variances of wavelet coefficients in the high frequency bands, are no longer good indicators for classes. We thus use the segmentation results obtained from previous scales as context, and classify based only on the closeness of average intensities.

The algorithm examines blocks on the boundaries of OOI and background because the boundary regions are most likely to be incorrectly classified. Consequently, the algorithm can be viewed as a multiscale edge refiner. At the boundary regions, a parent block may contain both classes. A sub-block which is incorrectly segmented as the class of its parent block is likely to have properties rather different from its parent block. In our case, the properties are the feature v and the average intensity. If a sub-block happens to have a closer average intensity to those of its neighboring blocks with the other class, and its own feature v is also closer to the other class, we switch its class.

After the multiscale refining of edges is completed, the segmented image is passed through a postprocessor, which removes small isolated regions and smoothes the boundaries.

5 Experimental Results

The system has been implemented by C on UNIX platforms. The experiments were performed on a Pentium Pro 300MHz LINUX workstations. The algorithm has achieved high accuracy when tested on more than 100 low DOF images, many with inhomogeneous foreground or background distractions. The segmentation results obtained were very close to the human partitioning of these images. Besides its high accuracy, the algorithm is very fast. An image of 768×512 pixels can be segmented within two seconds on a Pentium Pro 300MHz PC.

An example is shown in Figure 7 to illustrate how the progressive segmentation proceeds. Six scales are used starting with block size 32×32 and going to 1×1. The segmentation result at the scale of block size 2×2 is omitted due to limited space. Figure 8 shows the segmentation results for images in Figure 2(a), (b), (c).

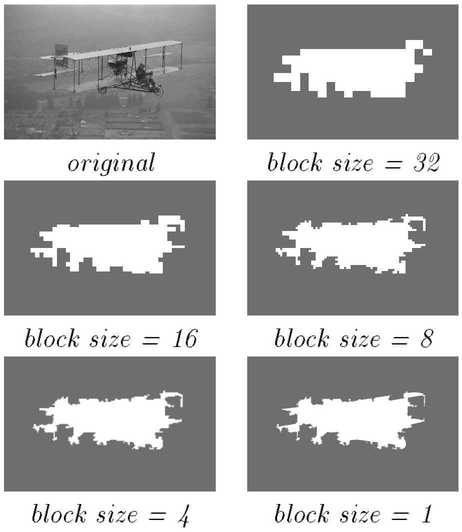


Figure 7. The sequence of multiscale refinement results.

We evaluate classification performance by *sensitivity*, *specificity*, and error rate. Sensitivity is defined as the ratio of the area of the OOI identified to the total area of OOI in the image. Specificity is defined as the ratio of the area of the background identified to the total area of the background in the image. For segmented images shown in Figure 8, the sensitivity, specificity, and error rate (P_e) are provided by Table 1. For a typical low DOF image, such as the jet plane image, the sensitivity and specificity are both above 95%. For a more difficult image, such as the fox image, the system demonstrates a high accuracy of 90% sensitivity and 80% specificity. We expect other edge-based or

Figure 8. First row: human segmentation. Second row: computer segmentation.

Image	Resolution	Sen.	Spec.	P_e
(a)	256×172	73.7%	97.5%	5.5%
(b)	768×512	90.7%	80.1%	16.1%
(c)	512×768	97.5%	96.4%	3.4%

Table 1. Segmentation results.

snake-based segmentation algorithms to perform worse on images with so much background distraction.

Our algorithm has the following limitations:

1. It cannot be applied to segment high DOF images.

2. It is not designed to segment those low DOF images for which some high-level human knowledge or image stereo is required in the human region determination process. Unlike many other edge-based segmentation algorithms, we rely on the sharp details of the OOI. However, if the OOI is highly smooth, the algorithm may fail.

3. It is designed to segment fully focused OOI. For some applications, the DOF may be so low that the OOI itself may include out-of-focus regions. The algorithm is not capable of segmenting the entire OOI in this case.

4. The performance of the algorithm is lower when the image resolution or the image quality is low.

6 Conclusions

In this paper, we have described a novel multiresolution context-dependent unsupervised image segmentation algorithm for low depth of field images. The algorithm has achieved high accuracy and high speed when tested on more than 100 low DOF images, many with inhomogeneous foreground or background distractions.

References

[1] A. Adams, *The Camera*, New York Graphic Society, Boston, 1980.

[2] Y. Boykov and D. Huttenlocher, A New Baysian Framework for Object Recognition, *Proc. DARPA IU Workshop*, Morgan Kauffmann, Monterey, 1998.

[3] T. Darell and K. Wohn, Depth From Focus Using a Pyramid Architecture, *Pattern Recognition Letters*, 11(12):787-96, Dec. 1990.

[4] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, 1992.

[5] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, Efficient and Effective Querying by Image Content, *J. of Intelligent Information Systems*, 3:231-262, 1994.

[6] R. G. Gonzalez and R. E. Woods, *Digital Image Processing*, Reading, MA, Addison-Wesley, 1992.

[7] A. Gupta and R. Jain, Visual Information Retrieval, *Communications of the ACM*, vol. 40, pp 69-79, 1997.

[8] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, N.J. : Prentice Hall, 1988.

[9] J. Li and R. M. Gray, Context Based Multiscale Classification of Images, *IEEE Int. Conf. Image Processing*, Chicago, Oct. 1998.

[10] E. Krotkov, *Active Computer Vision by Cooperative Focus and Stereo*, New York, Springer-Verlag, 1989.

[11] S. K. Nayar and Y. Nakagawa, Shape From Focus, *IEEE Trans. PAMI*, 16(8):824-31, Aug. 1994.

[12] M. Noguchi and S. K. Nayar, Microscopic Shape From Focus Using a Projected Illumination Pattern, *Mathematical and Computer Modelling*, Elsevier, 24(5-6):31-48, Sept. 1996.

[13] L. Stroebel (ed.), *Basic Photographic Materials and Processes*, Boston : Focal Press, c1990.

[14] M. Subbarao, T. Yuan, and J.-K. Tyan, Integration of Defocus and Focus Analysis with Stereo for 3D Shape Recovery, *Three-Dimensional Imaging and Laser-based Systems for Metrology and Inspection III*, Pittsburgh, PA, 14-15, Oct. 1997.

[15] J. Z. Wang, G. Wiederhold, O. Firschein, and X. W. Sha, Content-based Image Indexing and Searching Using Daubechies' Wavelets, *International Journal of Digital Libraries*, 1(4):311-328, Springer-Verlag, 1998.

[16] C. Yim and A. C. Bovik, Multiresolution 3-D Range Segmentation Using Focus Cues, *IEEE Tran. Image Processing*, 7(9):1283-1299, 1998.