Salient Region Detection and Segmentation

Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk

School of Computer and Communication Sciences (I&C),
Ecole Polytechnique Fédérale de Lausanne (EPFL),
{radhakrishna.achanta,francisco.estrada,patricia.wils,sabine.susstrunk}@epfl.ch
http://ivrg.epfl.ch/

Abstract. Detection of salient image regions is useful for applications like image segmentation, adaptive compression, and region-based image retrieval. In this paper we present a novel method to determine salient regions in images using low-level features of luminance and color. The method is fast, easy to implement and generates high quality saliency maps of the same size and resolution as the input image. We demonstrate the use of the algorithm in the segmentation of semantically meaningful whole objects from digital images.

Key words: Salient regions, low-level features, segmentation

1 Introduction

Identifying visually salient regions is useful in applications such as object based image retrieval, adaptive content delivery [11, 12], adaptive region-of-interest based image compression , and smart image resizing [2]. We identify salient regions as those regions of an image that are visually more conspicuous by virtue of their contrast with respect to surrounding regions. Similar definitions of saliency exist in literature where saliency in images is referred to as local contrast [9, 11].

Our method for finding salient regions uses a contrast determination filter that operates at various scales to generate saliency maps containing "saliency values" per pixel. Combined, these individual maps result in our final saliency map. We demonstrate the use of the final saliency map in segmenting whole objects with the aid of a relatively simple segmentation technique. The novelty of our approach lies in finding high quality saliency maps of the same size and resolution as the input image and their use in segmenting whole objects. The method is effective on a wide range of images including those of paintings, video frames, and images containing noise.

The paper is organized as follows. The relevant state of the art in salient region detection is presented in Section 2. Our algorithm for detection of salient regions and its use in segmenting salient objects is explained in Section 3. The parameters used in our algorithm, the results of saliency map generation, segmentation, and comparisons against the method of Itti et al. [9] are given in Section 4. Finally, in Section 5 conclusions are presented.

2 Approaches for Saliency Detection

The approaches for determining low-level saliency can be based on biological models or purely computational ones. Some approaches consider saliency over several scales while others operate on a single scale. In general, all methods use some means of determining local contrast of image regions with their surroundings using one or more of the features of color, intensity, and orientation. Usually, separate feature maps are created for each of the features used and then combined [8, 11, 6, 4] to obtain the final saliency map. A complete survey of all saliency detection and segmentation research is beyond the scope of this paper, here we discuss those approaches in saliency detection and saliency-based segmentation that are most relevant to our work.

Ma and Zhang [11] propose a local contrast-based method for generating saliency maps that operates at a single scale and is not based on any biological model. The input to this local contrast-based map is a resized and color quantized CIELuv image, sub-divided into pixel blocks. The saliency map is obtained from summing up differences of image pixels with their respective surrounding pixels in a small neighborhood. This framework extracts the points and regions of attention. A fuzzy-growing method then segments salient regions from the saliency map.

Hu et al. [6] create saliency maps by thresholding the color, intensity, and orientation maps using histogram entropy thresholding analysis instead of a scale space approach. They then use a spatial compactness measure, computed as the area of the convex hull encompassing the salient region, and saliency density, which is a function of the magnitudes of saliency values in the saliency feature maps, to weigh the individual saliency maps before combining them.

Itti et al. [9] have built a computational model of saliency-based spatial attention derived from a biologically plausible architecture. They compute saliency maps for features of luminance, color, and orientation at different scales that aggregate and combine information about each location in an image and feed into a combined saliency map in a bottom-up manner. The saliency maps produced by Itti's approach have been used by other researchers for applications like adapting images on small devices [3] and unsupervised object segmentation [5, 10].

Segmentation using Itti's saliency maps (a 480x320 pixel image generates a saliency map of size 30x20 pixels) or any other sub-sampled saliency map from a different method requires complex approaches. For instance, a Markov random field model is used to integrate the seed values from the saliency map along with low-level features of color, texture, and edges to grow the salient object regions [5]. Ko and Nam [10], on the other hand, use a Support Vector Machine trained on the features of image segments to select the salient regions of interest from the image, which are then clustered to extract the salient objects. We show that using our saliency maps, salient object segmentation is possible without needing such complex segmentation algorithms.

Recently, Frintrop et al. [4] used integral images [14] in VOCUS (Visual Object Detection with a Computational Attention System) to speed up computation of center-surround differences for finding salient regions using separate

feature maps of color, intensity, and orientation. Although they obtain better resolution saliency maps as compared to Itti's method, they resize the feature saliency maps to a lower scale, thereby losing resolution. We use integral images in our approach but we resize the filter at each scale instead of the image and thus maintain the same resolution as the original image at all scales.

3 Salient region detection and segmentation

This section presents details of our approach for saliency determination and its use in segmenting whole objects. An overview of the complete algorithm is presented in Figure 1. Using the saliency calculation method described later, saliency maps are created at different scales. These maps are added pixel-wise to get the final saliency maps. The input image is then over-segmented and the segments whose average saliency exceeds a certain threshold are chosen.

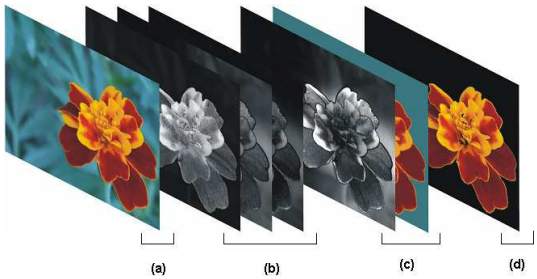

Fig. 1. Overview of the process of finding salient regions. (a) Input image. (b) Saliency maps at different scales are computed, added pixel-wise, and normalized to get the final saliency map. (c) The final saliency map and the segmented image. (d) The output image containing the salient object that is made of only those segments that have an average saliency value greater than the threshold T (given in Section 3.1).

3.1 Saliency calculation

In our work, saliency is determined as the local contrast of an image region with respect to its neighborhood at various scales. This is evaluated as the distance between the average feature vector of the pixels of an image sub-region with the average feature vector of the pixels of its neighborhood. This allows obtaining a combined feature map at a given scale by using feature vectors for each pixel, instead of combining separate saliency maps for scalar values of each feature. At a given scale, the contrast based saliency value $c_{i,j}$ for a pixel at position (i, j) in the image is determined as the distance D between the average vectors of pixel

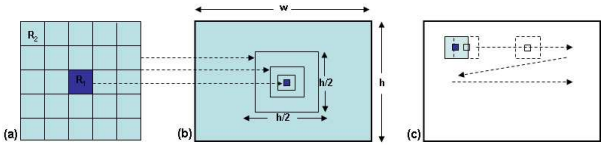

Fig. 2. (a) Contrast detection filter showing inner square region R_1 and outer square region R_2. (b) The width of R_1 remains constant while that of R_2 ranges according to Equation 3 by halving it for each new scale. (c) Filtering the image at one of the scales in a raster scan fashion.

features of the inner region R_1 and that of the outer region R_2 (Figure 2) as:

$$c_{i,j} = D\left[\left(\frac{1}{N_1}\sum_{p=1}^{N_1}\mathbf{v}_p\right), \left(\frac{1}{N_2}\sum_{q=1}^{N_2}\mathbf{v}_q\right)\right] \tag{1}$$

where $N1$ and N_2 are the number of pixels in R_1 and R_2 respectively, and $\mathbf{v}$ is the vector of feature elements corresponding to a pixel. The distance D is a Euclidean distance if $\mathbf{v}$ is a vector of uncorrelated feature elements, and it is a Mahalanobis distance (or any other suitable distance measure) if the elements of the vector are correlated. In this work, we use the $CIELab$ color space [7], assuming sRGB images, to generate feature vectors for color and luminance. Since perceptual differences in $CIELab$ color space are approximately Euclidian, D in Equation 1 is:

$$c_{i,j} = \|\mathbf{v}_1 - \mathbf{v}_2\| \tag{2}$$

where $\mathbf{v}_1 = [L_1, a_1, b_1]^T$ and $\mathbf{v}_2 = [L_2, a_2, b_2]^T$ are the average vectors for regions R_1 and R_2, respectively. Since only average feature vector values of R_1 and R_2 need to be found, we use the integral image approach as used in [14] for computational efficiency. A change in scale is affected by scaling the region R_2 instead of scaling the image. Scaling the filter instead of the image allows the generation of saliency maps of the same size and resolution as the input image. Region R_1 is usually chosen to be one pixel. If the image is noisy (for instance if high ISO values are used when capturing images, as can often be determined with the help of Exif data (Exchangeable File Information Format [1]) then R_1 can be a small region of $N \times N$ pixels (in Figure 5(f) N is 9).

For an image of width w pixels and height h pixels, the width of region R_2, namely w_{R_2} is varied as:

$$\frac{w}{2} \geq (w_{R_2}) \geq \frac{w}{8} \tag{3}$$

assuming w to be smaller than h (else we choose h to decide the dimensions of R_2). This is based on the observation that the largest size of R_2 and the smaller ones (smaller than $w/8$) are of less use in finding salient regions (see Figure 3). The former might highlight non-salient regions as salient, while the latter are basically edge detectors. So for each image, filtering is performed at three

Fig. 3. From left to right, original image followed by filtered images. Filtering is done using R_1 of size one pixel and varying width of R_2. When R_2 has the maximum width, certain non salient parts are also highlighted (the ground for instance). It is the saliency maps at the intermediate scales that consistently highlight salient regions. The last three images on the right mainly show edges.

different scales (according to Eq. 3) and the final saliency map is determined as a sum of saliency values across the scales S:

$$m_{i,j} = \sum_{S} c_{i,j} \tag{4}$$

$\forall\, i \in [1, w], j \in [1, h]$ where $m_{i,j}$ is an element of the combined saliency map $\mathbf{M}$ obtained by point-wise summation of saliency values across the scales.

3.2 Whole Object Segmentation using Saliency Maps

The image is over-segmented using a simple *K-means* algorithm. The K seeds for the *K-means* segmentation are automatically determined using the hill-climbing algorithm [13] in the three-dimensional $CIELab$ histogram of the image. The

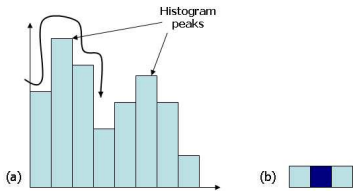

Fig. 4. (a) Finding peaks in a histogram using a search window like (b) for a one dimensional histogram.

hill-climbing algorithm can be seen as a search window being run across the space of the d-dimensional histogram to find the largest bin within that window. Figure 4 explains the algorithm for a one-dimensional case. Since the $CIELab$ feature space is three-dimensional, each bin in the color histogram has $3^d - 1 = 26$ neighbors where d is the number of dimensions of the feature space. The number

of peaks obtained indicates the value of K, and the values of these bins form the initial seeds.

Since *K-means* algorithm clusters pixels in the $CIELab$ feature space, an 8-neighbor connected-components algorithm is run to connect pixels of each cluster spatially. Once the segmented regions r_k for $k = 1, 2...K$ are found, the average saliency value V per segmented region is calculated by adding up values in the final saliency map $\mathbf{M}$ corresponding to pixels in the segmented image:

$$V_k = \frac{1}{|r_k|} \sum_{i,j \in r_k} m_{i,j} \tag{5}$$

where $|r_k|$ is the size of the segmented region in pixels. A simple threshold based method can be used wherein the segments having average saliency value greater than a certain threshold T are retained while the rest are discarded. This results in an output containing only those segments that constitute the salient object.

4 Experiments and Results

Experiments were performed on images from the Berkely database and from flickr$^{TM,\ 1}$. The saliency maps for Itti's model[2] were generated using iLAB Neuromorphic Vision Toolkit[3]. The results of salient region segmentation from our method[4] are compared with those from Itti's model for the same input image and same segmentation algorithm. For segmentation, a window size of $3 \times 3 \times 3$ is used for the hill-climbing search on a $16 \times 16 \times 16$ bin $CIELab$ histogram. The average saliency threshold used for selecting segments T is set at 25 (about 10% of the maximum possible average saliency in the normalized final saliency map) based on observations on about 200 images. This threshold is not too sensitive and can be varied by 10% of its value without affecting the segmentation results. The results[5] in Figures 6 and 7 show that salient pixels using our computational method correspond closely to those using Itti's method, which is based on a biological model. In addition, because of the high resolution of the saliency maps, the entire salient region is clearly highlighted (Figures 6 and 7, column 3). This facilitates a clean segmentation of semantically more meaningful whole objects without having to use an overly complex segmentation algorithm.

We compared speed of salient map generation of our proposed method against that of Itti's method. The results are shown in Table 1. Our algorithm is at least five times faster in generation of saliency maps for a given image size. Although both algorithms have roughly have a complexity of $O(n)$ (which is also evident

[1] http://www.flickr.com/

[2] Since Itti's model generates very small saliency maps relative to the original input image, in Figures 6 and 7 these images are shown up-scaled.

[3] http://ilab.usc.edu/toolkit/

[4] http://ivrg.epfl.ch/~achanta/SalientRegionDetection/SalientRegionDetection.html

[5] The saliency maps from Itti's method as well as our method shown in the results are contrast-stretched for better print quality.

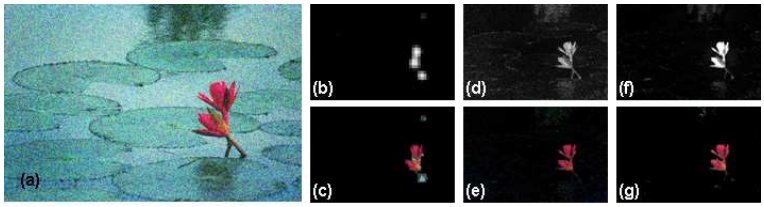

Fig. 5. (a) Original image with 5db gaussian noise. (b) Itti's saliency map. (c) Segmentation result using map (b). (d) Saliency map with our method using R_1 of size 1×1. (e) Segmentation result using map (d). (f) Saliency map with our method using R_1 of size 9×9. (g) Segmentation result using map (f).

from the speeds vs. image size values in Table 1), there is a lot more processing taking place in Itti' method, where apart from color and luminance maps several orientation maps are also created. As opposed to this only three maps created by our method for the features of color and luminance treated as one vector value. Itti's method computes center-surround differences by performing subtraction between Gaussian pyramid levels. This speedup results in a loss of resolution. Our method instead changes the size of the filter at each scale through the use of integral images, which achieves even greater speed without lowering the resolution.

Table 1. Table comparing time (in seconds) required to find salient regions for different sizes of input images. The two algorithms were run on an Intel Dual Core 2.26 GHz machine with 1GB RAM.

Algorithm used	**320x240**	**640x480**	**800x600**	**1024x768**
Itti-Koch Method	0.75	2.54	4.40	7.50
Our algorithm for saliency	0.12	0.46	0.68	1.29

In cases when the salient object occupies a large part of the image or if certain parts of the salient objects do not show sufficient contrast w.r.t their surroundings (eg. the back of the deer in Figure 7), the salient object segmentation is not satisfactory. At times there are some holes left in the salient objects or some extra segments are present in the final result (eg. spots of the Jaguar in Figure 7). These can be handled in a post-processing step. In the experiments done with noisy images it was observed that (see Figure 5), for moderate amounts of noise (less than 1dB), one pixel size for R_1 suffices. The size of R_1 can be increased in the presence greater amount of noise for better salient region detection.

5 Conclusions

We presented a novel method of finding salient regions in images, using low level features of color and luminance, which is easy to implement, noise tolerant, and

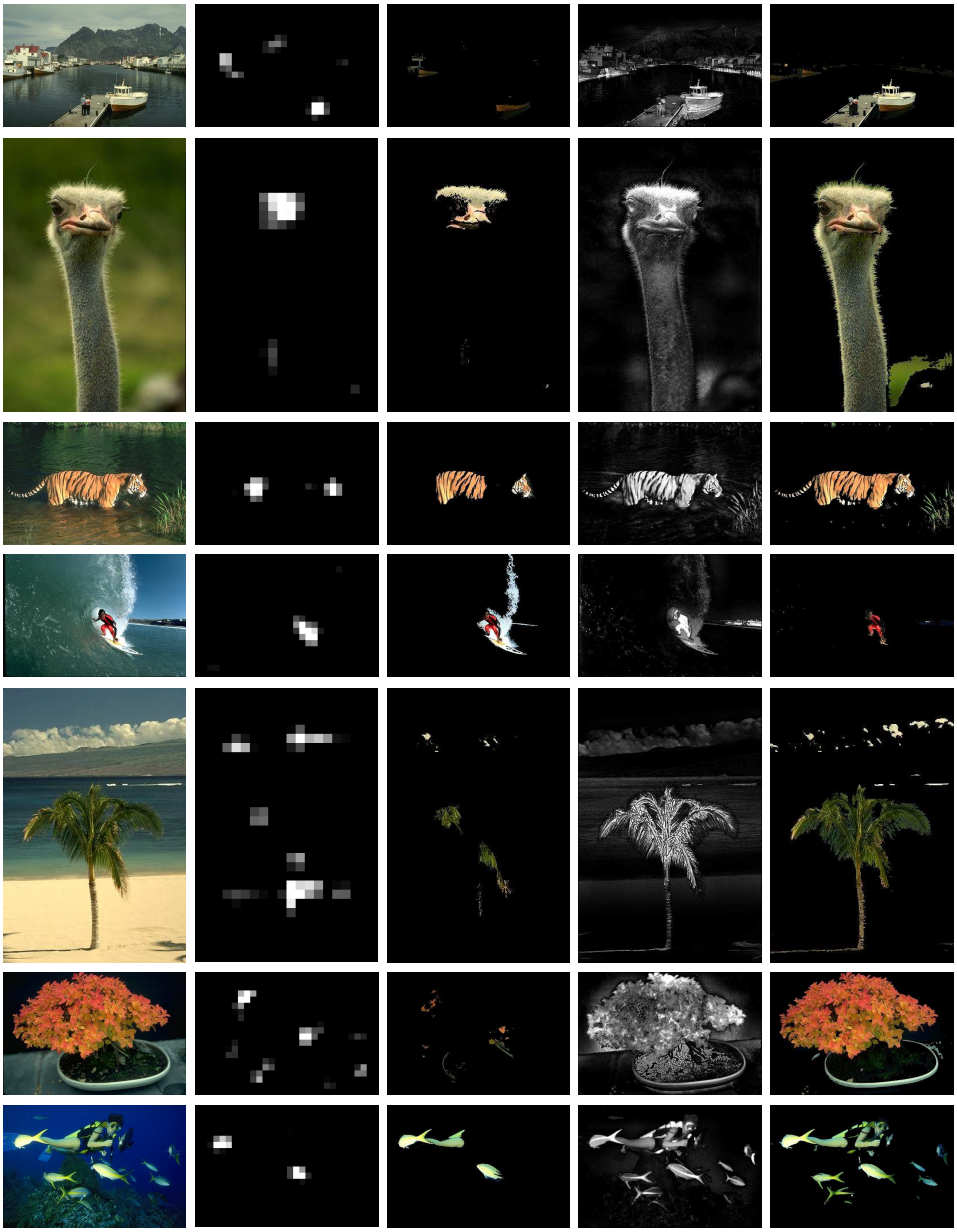

Fig. 6. Visual attention region detection results on images from the Berkeley database. From left to right: Original image, Itti's saliency map, segmentation using Itti's map, saliency map using our method, and segmentation using our saliency map. Note that the regions of saliency in Itti's maps and our maps are often the same, however, in our maps, the detail is much greater and the regions are well defined.

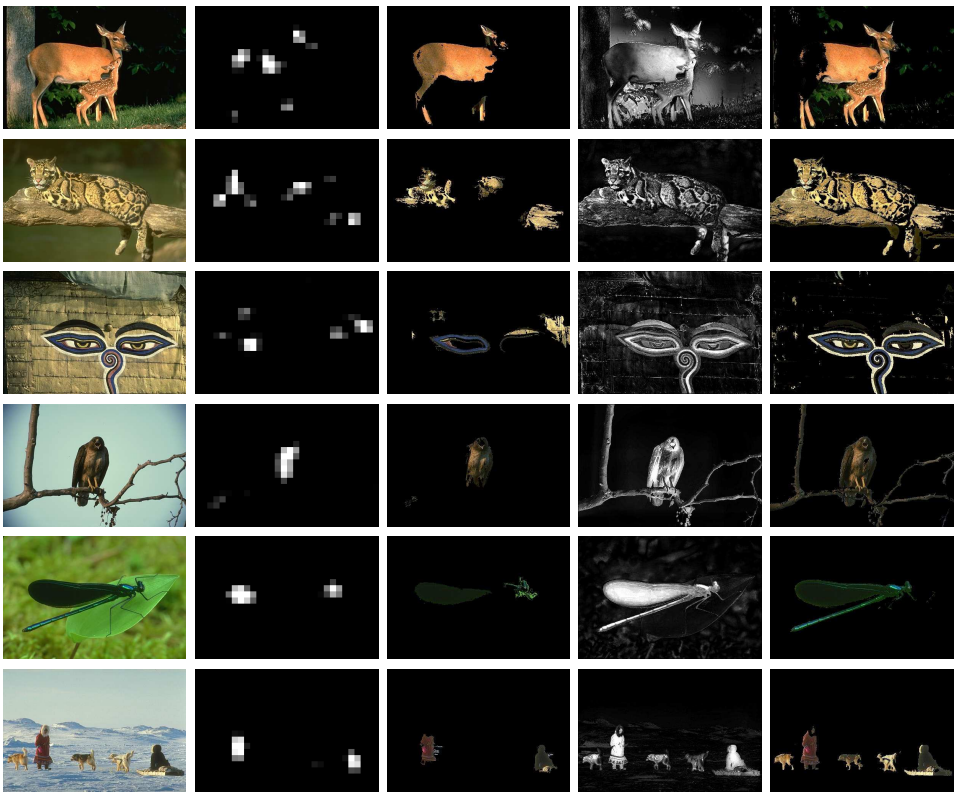

Fig. 7. Visual attention region detection results on images from the Berkeley database. From left to right: Original image, Itti's saliency map, segmentation using Itti's map, saliency map using our method, and segmentation using our saliency map.

fast enough to be useful for real time applications. It generates saliency maps at the same resolution as the input image. We demonstrated the effectiveness of the method in detecting and segmenting salient regions in a wide range of images. The approach is at least five times as fast as a prominent approach to finding saliency maps and generates high resolution saliency maps that allow better salient object segmentation.

6 Acknowledgements

This work is supported by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322, and the European Commission under contract FP6-027026 (K-Space, the

European Network of Excellence in Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content).

References

1. Digital still camera image file format standard (exchangeable image file format for digital still cameras: Exif) Version 2.1, Specification by JEITA, June 1998.
2. S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, 26(3):10, July 2007.
3. L. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H. Zhou. A visual attention model for adapting images on small displays. *ACM Transactions on Multimedia Systems*, 9:353–364, November 2003.
4. S. Frintrop, M. Klodt, and E. Rome. A real-time visual attention system using integral images. In *International Conference on Computer Vision Systems (ICVS'07)*, March 2007.
5. J. Han, K. Ngan, M. Li, and H. Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(1):141–145, January 2006.
6. Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, and D. Rajan. Salient region detection using weighted feature maps based on the human visual attention model. *Springer Lecture Notes in Computer Science*, 3332(2):993–1000, October 2004.
7. R. W. G. Hunt. *Measuring Color*. Fountain Press, 1998.
8. L. Itti and C. Koch. Comparison of feature combination strategies for saliency-based visual attention systems. In *SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, pages 473–482., May 1999.
9. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.
10. B. C. Ko and J.-Y. Nam. Object-of-interest image segmentation based on human attention and semantic region clustering. *Journal of Optical Society of America A*, 23(10):2462–2470, October 2006.
11. Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, pages 374–381, November 2003.
12. V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch. Automatic image retargeting. In *Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia (MUM'05)*, pages 59–68, October 2005.
13. T.Ohashi, Z.Aghbari, and A.Makinouchi. Hill-climbing algorithm for efficient color-based image segmentation. In *IASTED International Conference On Signal Processing, Pattern Recognition, and Applications (SPPRA 2003)*, June 2003.
14. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'01)*, 1:511–518, December 2001.