AN ATTENTION-BASED METHOD FOR EXTRACTING SALIENT REGIONS OF INTEREST FROM STEREO IMAGES

Oge Marques, Liam M. Mayron, Daniel Socek
Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL - 33431, USA

Gustavo B. Borba, Humberto R. Gamba
CPGEI, Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro 3165, Curitiba-PR, Brazil

Keywords: Image segmentation, stereo vision, visual attention.

Abstract: A fundamental problem in computer vision is caused by the projection of a three-dimensional world onto one or more two-dimensional planes. As a result, methods for extracting regions of interest (ROIs) have certain limitations that cannot be overcome with traditional techniques that only utilize a single projection of the image. For example, while it is difficult to distinguished two overlapping, homogeneous regions with a single intensity or color image, depth information can usually easily be used to separate the regions. In this paper we present an extension to an existing saliency-based ROI extraction method. By adding depth information to the existing method many previously difficult scenarios can now be handled. Experimental results show consistently improved ROI segmentation.

1 INTRODUCTION

Extracting regions of interest (ROIs) from digital images represents one of the fundamental tasks in computer vision. The problem of extracting ROIs from digital images of natural scenes is often exacerbated by the loss of information caused by a two-dimensional projection of the three-dimensional real world. Consequently, most methods have difficulty distinguishing homogeneous overlapping regions caused by partial occlusion, or separating regions belonging to the background from those belonging to the foreground.

In this paper we present a method for extracting salient regions of interest from stereo images. Our approach represents an extension to an existing attention-based ROI extraction method proposed in (Marques et al., 2007), which relies on two complementary computational models of human visual attention, (Itti et al., 1998) and (Stentiford, 2003). These models provide important cues about the location of the most salient ROIs within an image. By incorporating the estimated depth information obtained from left and right stereo images, our method can successfully cope with the aforementioned extraction problems, resulting in a more robust and versatile method.

2 THE PROPOSED METHOD

The proposed method combines a saliency-based ROI extraction method with depth map information. This Section presents background information on both topics and outlines the proposed solution.

Much of the visual information our eyes sense is discarded. Instead,

Marques O., M. Mayron L., Socek D., B. Borba G. and R. Gamba H. (2007).
AN ATTENTION-BASED METHOD FOR EXTRACTING SALIENT REGIONS OF INTEREST FROM STEREO IMAGES.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 294-297

tiford, 2003) through a series of morphological operations. The model produces one or more extracted regions of interest.

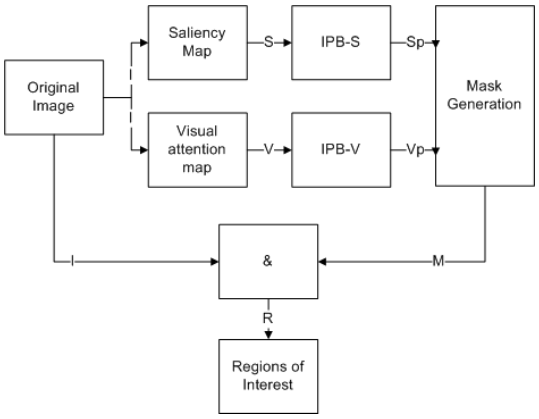


Figure 1: General block diagram of the 2D ROI extraction method.

Figure 1 shows an overview of the 2D ROI extraction method. The saliency map (Itti-Koch) (S) and visual attention map (Stentiford) (V) are generated from the original image. Post-processing is performed independently on each in order to remove stray points and prune potential regions. Then, the remaining points in the processed saliency map are used to target regions of interest that remain on the visual attention map. The result is a mask (M) that can be used to extract the regions of interest (R) from the original image. This process is detailed in (Marques et al., 2007).

There are certain cases where the previous method does not work. When objects are occluded or overlapping they may appear as a single region when inspecting a single 2D projection of the view. Only with a separate view can enough information of the original 3D scene be reconstructed to determine the relative depth of the occluding objects. Convesely, relying only on depth information is also not enough to properly determine a region of interest. A bright poster on a flat wall, for example, would be ignored if only depth information were used, as it rests on the same plane as the wall. As a result, we propose a combination of both methods, mitigating the weaknesses of each.

In (Birchfield and Tomasi, 1999), the authors proposed fast and effective algorithm for depth estimation from stereo images. Unlike other similar approaches, such as (Cox et al., 1996), the approach of Birchfield and Tomasi achieves optimal performance mainly by avoiding subpixel resolution with a measure that is insensitive to image sampling. The depth estimation phase of our method relies on this computational approach.

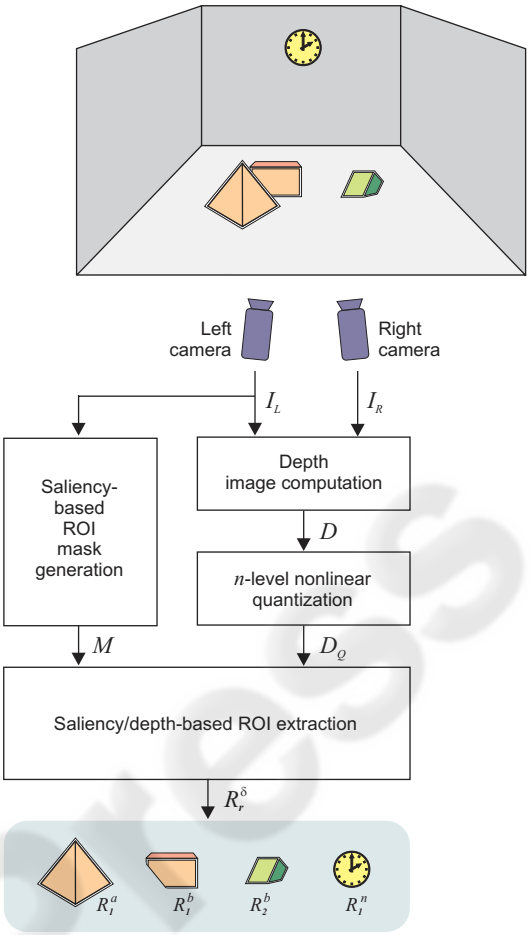


Figure 2: General block diagram of the 3D ROI extraction.

According to Figure 2, the scene is first acquired by two properly-positioned and adjusted cameras, so that the scanlines are the epipolar lines. The left and right stereo images, I_L and I_R are processed by Birchfield-Tomasi disparity estimation algorithm. The output disparity map D is then *nonlinearly quantized* within n levels, resulting in output image D_Q. The left channel image, I_L, is also processed by the existing 2D saliency-based ROI segmentation algorithm that produces a binary mask M corresponding to the salient regions of the image (Figure 1). In the last stage of the algorithm, M and D_Q are submitted to the *saliency/depth-based ROI extraction* block, which combines both images in order to segment the ROIs (R_r^δ) and label them according to their respective depths in the real scene. δ is the quantized depth, with $\delta \in \{a, \ldots, n\}$ and r is an ROI at a depth δ.

In the example shown in Figure 2, the objects (ROIs) belong to either foreground, middle, or background. In the output at the bottom of the figure the pyramid within the foreground plane is labeled with R_1^a, the partially occluded parallelepiped and the green

solid, at the same middle plane, are labeled with R_1^b and R_2^b. Finally, the clock in the background is labeled with R_1^n.

Under normal conditions depth images are relatively efficient in discriminating objects at the frontal planes of the scene but they generally do not have sufficient resolution to capture flat objects in the background or even common objects on a distant plane. On the other hand, a saliency-based ROI identification algorithm can capture such objects, but they do not account for relative object depth within the scene. The objective is to combine the information provided by both salient regions and depth cues to improve ROI extraction.

In Figure 2, a purely saliency-driven ROI extraction algorithm tends to identify both light-orange objects as a single region. However, using depth information, it is possible to divide this region, discriminating the two objects. Another benefit of this approach is the possibility of extracting objects such as the watch in the background of Figure 2. While algorithms for depth estimation are not able to discriminate the watch plane from the wall plane (their depth is too similar), a saliency-driven ROI extraction can segment that object. Using only depth images the watch would not be captured.

3 COMPONENTS

The following is a description of the system components from Figure 2.

Depth Images

The disparity maps generated by the Birchfield-Tomasi method are represented as 256-level grayscale images. Darker (lower) values indicate further distances, and vice versa. In particular, purely black values denote the background plane.

Nonlinear Quantization

An n-level ($L_1, \ldots, L_n$) quantization is obtained and applied to the disparity map according to Equation 1. Level L_1 identifies the depth closest to the cameras and level L_n denotes the depth farthest depth from camera (the background).

$$D_Q(x,y) = \begin{cases} L_n & \text{if } D(x,y) = [0 \;\; T_1), \\ L_{n-1} & \text{if } D(x,y) = [T_1 \;\; T_2), \\ \vdots & \\ L_1 & \text{if } D(x,y) = [T_{n-1} \;\; 255]. \end{cases} \quad (1)$$

where T_i are the selected threshold values.

Saliency-based Roi Mask

Salient regions of interest are extracted from the left image using the method described in (Marques et al., 2007). This method was modified in the original saliency-driven ROI extraction algorithm to refine some of the thresholds used to determine relative object size.

Roi Extraction

The ROI extraction stage combines images M and D_Q. Its goal is to segment and label the ROIs according to their depths in the real scene. First, an *AND* operation between grayscale image D_Q and mask M is performed, originating a grayscale $\mathcal{D}$ image. This image is then used to perform a *depth decomposition* according to Equation 2.

$$\mathcal{D}^{\delta} = \begin{cases} 1 & \text{if } \mathcal{D}(x,y) = L_{\delta}, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\mathcal{D}^{\delta}$ are the decomposed depths binary images. δ is the depth, with $\delta \in \{1, \ldots, n\}$.

After that, ROIs can be effectively extracted. First, decomposed depth image $\mathcal{D}^1$ is submitted to a set of morphological operations, denoted by $m(\cdot)$, in Equation 3.

$$R^1 = m(\mathcal{D}^1) \quad (3)$$

R^1 is a binary image where the white regions corresponds to ROIs into depth 1, that is, those that are closest to the camera. Function $m(\cdot)$ performs the following sequence: closing, region filling, pruning, and small blobs elimination.

The remaining R^{δ} for each decomposed depth are sequentially computed, from $\delta = 2$ to $\delta = n$, according to Equation 4

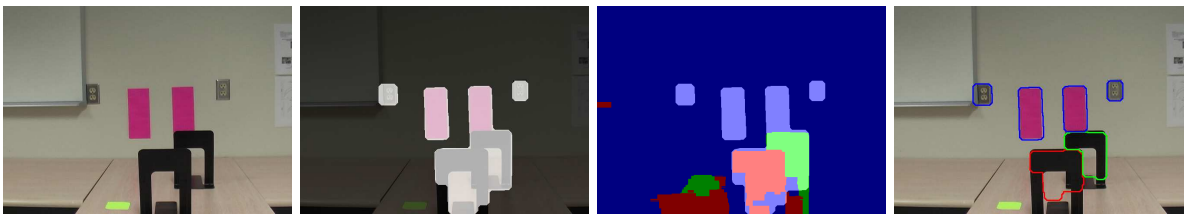
$$R^{\delta} = m\left(\mathcal{D}^{\delta} \bigcap \left[\bigcup_{k=1}^{\delta-1} R^k\right]^c\right) \quad (4)$$

where $[\cdot]^c$ means the complement operation. Note that the computation of a deeper R^{δ} takes into account the depths before it. This operations gives preference to closer regions of interest over the further ones.

Each image R^{δ} can have a set of ROIs, denoted by:

$$\{R^{\delta}\} = \{R_1^{\delta}, \ldots, R_r^{\delta}\} \quad (5)$$

where r is the number of ROIs in the depth δ, with $r \geq 0$.



(a) (b) (c) (d)

Figure 3: Results for occluding salient objects in the foreground and distracting salient objects in the background: (a) original left stereo image; (b) highlighted ROI using saliency-based mask M; (c) highlighted pseudo-colored nonlinearly quantized depth image (D_Q) using binary mask M. From the closest to the deepest plane: red, green, blue; and (d) final ROIs highlighted in the actual image.

4 EXPERIMENTAL RESULTS

We illustrate the performance of our algorithm with a setting (Figure 3) consisting of stereo images captured in laboratory environment with two aligned identical cameras fixed on a professional stereo stand. This and other stereo image pairs along with the experimental results are currently posted at `http://mlab.fau.edu/stereo/roi3d.zip`.

In our experiments, a 3-level (L_1, L_2, L_3) quantization was used, according to Equation 1, while the threshold values were obtained empirically: $T_1 = 11$ and $T_2 = 23$.

Figure 3 shows a case in which occluding salient objects in the foreground and distracting objects in the background are segmented. Note that there is a bright yellow distracter in the foreground that is not perceived as such by the algorithm, resulting in a false negative. It can be observed that while the 2D ROI extraction fails to discriminate between two foreground objects and fails to identify background objects as such, our proposed algorithm successfully discriminates between the two foreground ROIs and identifies all background ROIs.

5 CONCLUSIONS

Object and region segmentation from 2D data is not always a straightforward task. In particular, it can be impossible to segment occluded object because of the depth information that is lost. In this work we extended a previously proposed method for 2D region of interest extraction with depth information. A disparity map was generated from two views using the method proposed by Birchfield-Tomasi (Birchfield and Tomasi, 1999). Using this depth information we were able to differentiate occluding regions of interest. Our experiments demonstrate the promise of this approach but stress the need for nonlinear quantization thresholds of the disparity map for successful results. We are continuing work on this approach by creating a method of automatically determining these quantization thresholds and extending it to a variety of applications. We are currently obtaining quantitative results to further validate our method.

ACKNOWLEDGEMENTS

This research was partially sponsored by UOL (www.uol.com.br), through its *UOL Bolsa Pesquisa* program, process number 200503312101a.

REFERENCES

Birchfield, S. and Tomasi, C. (1999). Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293.

Cox, I., Hingorani, S., Rao, S., and Maggs, B. (1996). A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20(11):1254–1259.

Marques, O., Mayron, L. M., Borba, G. B., and Gamba, H. R. (2007). An attention-driven model for grouping similar images with image retrieval applications. *EURASIP Journal on Applied Signal Processing*.

Stentiford, F. (2003). An attention based similarity measure with application to content-based information retrieval. In *Proceedings of the Storage and Retrieval for Media Databases Conference, SPIE Electronic Imaging*, Santa Clara, CA.

Styles, E. A. (2005). *Attention, Perception, and Memory: An Integrated Introduction*. Taylor & Francis Routledge, New York, NY.