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

Keywords: Image segmentation, stereo vision, visual attention.

Abstract: The fundamental problem of computer vision is caused by the translation 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.

The paper is organized as follows. We present background information in Section 2. The proposed approach for ROI extraction from stereo images is described in Section 3, while a more detailed analysis of its main components is presented in Section 4. The experimental results are given in Section 5. Section 6 concludes this paper.

2 BACKGROUND

This section provides background information on two main components of the proposed solution: computation models of visual attention and depth estimation techniques.

2.1 Computational Models of Visual Attention

Much of the visual information our eyes sense is discarded. Instead, our brain prioritizes what points in a scene we focus our attention on. The results is a series of fixations and saccades known as *scanpaths* (Noton and Stark, 1971).

There are two ways attention manifests itself;

bottom-up and top-down. The former is rapid, involuntary, and in reaction to the stimulus which is presented (Styles, 2005). Only later does top-down attention take place. It is motivated by our past knowledge and memories (Styles, 2005). Both play a role in how our attention is ultimately guided, but to what extent remains unclear.

However, it is clear that top-down attention is a complex process, whereas bottom-up attention is far more consistent, simpler, and more well-defined. Hence, the computational modeling of bottom-up processes of visual attention has been most successful to date. For a review of existing computational models of visual attention please refer to (Itti and Koch, 2001).

2.2 Depth Estimation from Stereo Images

Given a pair of stereo images, the *correspondence problem* refers to finding the *match sequence* for each left and right image scanline. The *match* refers to an ordered pair (x,y), where x and y are the positions in same scanlines of left and right stereo pair, respectively, such that the pixel values corresponding to these positions, $I_L(x)$ and $I_R(y)$, represent images of the same scene point. Here, it is assumed that the stereo images are properly aligned so that the scanlines are the epipolar lines. Unmatched pixels are labelled as *occluded*, and adjacent occluded pixels bounded by non-occluded pixels are called an *occlusion*.

The *disparity* $\Delta(x)$ of a pixel position x in the left scanline that matches the pixel y in the right scanline is defines as the difference $x-y$, while the disparities of the pixels in an occlusion are assigned the farther of the two bounding regions. Approaches to the stereo correspondence problem construct the so called *disparity map*, which is also often called the *depth map* or the *depth estimation* since it describes the discrete estimation of third spatial dimension.

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) (Geiger et al., 1995) (Intille and Bobick, 186), 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. Details of Birchfield-Tomasi algorithm can be found in (Birchfield and Tomasi, 1999), while (Birchfield and Tomasi, 1998) contains a detailed description of the proposed measure.

3 THE PROPOSED METHOD

3.1 An attention-driven model for extracting salient regions of interest

Recent studies reveal that several biologically-motivated models can be successfully applied to ROI extraction, targeting applications such as image exploration (Machrouh and Tarroux, 2005), target detection (Itti et al., 2001), and content-based image retrieval (Stentiford, 2003) and (Marques et al., 2007).

Our previous work (Marques et al., 2007) demonstrated a method of extracting regions of interest based on their saliency. It integrated the Itti-Koch model of visual attention (Itti et al., 1998) and that from Stentiford (Stentiford, 2003) through a series of morphological operations. The model produces one or more extracted regions of interest.

The Itti-Koch model of visual attention is bottom-up. The model generates a map of the most salient points in an image derived from color, intensity, orientation, motion, and possibly other features. Like the Itti-Koch model, the Stentiford model is also bottom-up. It suppresses areas of the image where patterns that are repeated elsewhere occur. As a result plain surfaces are suppressed while unique regions are brought to prominence. Regions are marked as high interest if they possess features not frequently present elsewhere in the image. The visual attention map generated by the Stentiford model tends to identify larger and smoother salient regions of an image as opposed to the more focused peaks in Itti-Koch's saliency map. Unfortunately, the tendency of the Stentiford model to mark large regions can lead to poor results if these regions are not salient. Itti's model is much better in this regard. By identifying the strengths and weaknesses of each model we were able to construct our method for extracting regions of interest from 2D images.

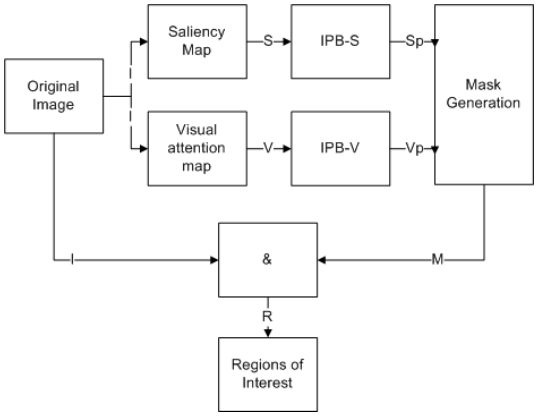


Figure 1: General block diagram of the 2D ROI extraction method.

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.

3.2 Overview

The proposed solution to ROI extraction from stereo images is summarized in a block diagram within Figure 2.

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

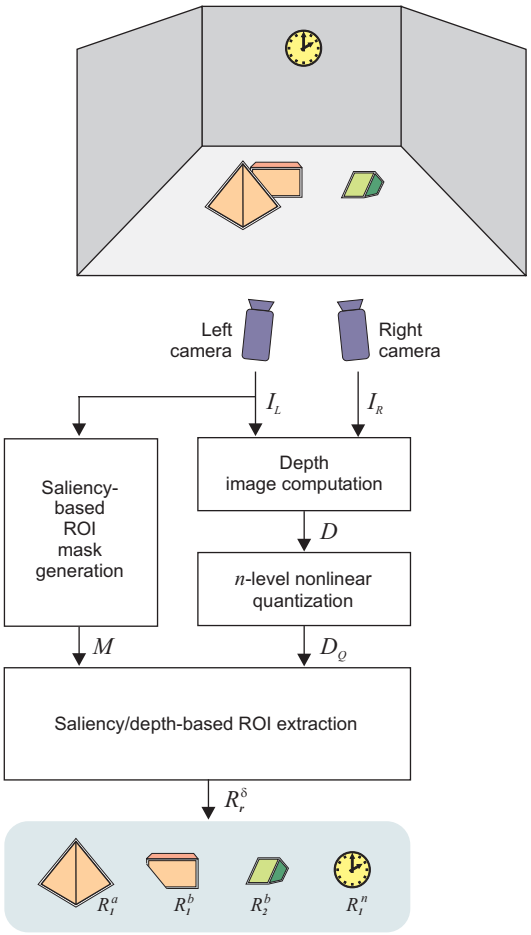


Figure 2: General block diagram of the 3D ROI extraction.

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 discrim-

inate 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.

4 COMPONENTS

The following section presents a detailed description of the system components from the block diagram depicted in Figure 2.

4.1 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.

4.2 Nonlinear quantization

An n-level (L_1, ..., 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.

4.3 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.

4.4 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 $\delta \in \{1,\ldots,n\}$.

ROIs are 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:

1. Closing: fills small gaps within the white pixels regions. Implemented using the "imclose()" function in MATLAB.
2. Region filling: flood-fills enclosed black pixels regions. Accomplished using the "imfill()" MATLAB function.
3. Pruning: performs a morphological opening and keeps only the largest remaining connected component, thereby eliminating smaller (undesired) branches.
4. Small blobs elimination: removes unconnected regions with area smaller then a fixed number of pixels.

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 $r \geq 0$

5 EXPERIMENTAL RESULTS

5.1 Method

In order to illustrate the performance of our algorithm four different experiments with different settings were

performed. The four different settings are depicted in Figures 3, 4, 5, and 6. The stereo images used in our experiments were captured in laboratory environment with two aligned identical cameras fixed on a professional stereo stand. The 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 6, while the threshold values were obtained empirically.

$$D_Q(x,y) = \begin{cases} L_3 & \text{if } D(x,y) = [0\ 11), \\ L_2 & \text{if } D(x,y) = [11\ 23), \\ L_1 & \text{if } D(x,y) = [23\ 255]. \end{cases} \quad (6)$$

The method was implemented using MATLAB code and we employed an implementation of Birchfield-Tomasi depth estimation by John Abd-El-Malek that presently can be found at `http://vision.stanford.edu/~birch/p2p/`. The maximum disparity for our set of stereo images was set to $\Delta = 30$.

5.2 Discussion

Figure 3 shows the easiest case in which two non-occluding salient objects in the foreground and two fixed salient objects in the background are properly segmented. Figure 4 shows the case where two distracting salient objects in the background are also segmented properly thanks to the disparity information that allows the boundary between those distracters and the foreground objects to be determined. Figure 5 shows a case where occluding salient objects in the foreground are properly segmented with disparity information. Finally, Figure 6 shows the most challenging combination in which occluding salient objects in the foreground and distracting objects in the background are segmented. Note that in Figures 5 and 6 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.

6 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 and by the Office of Naval Research (ONR) under the Center for Coastline Security Technology grant N00014-05-C-0031.

REFERENCES

Birchfield, S. and Tomasi, C. (1998). A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406.

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.

Geiger, D., Landendorf, B., and Yuille, A. (1995). Occlusions and binocular stereo. *International Journal of Computer Vision*, 14(3):211–226.

Intille, S. and Bobick, A. (1994, pages = 179–186). Disparity-space images and large occlusion stereo. In *Proceedings of the 3rd European Conference on Computer Vision*.

Itti, L., Gold, C., and Koch, C. (2001). Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793.

Itti, L. and Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

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.

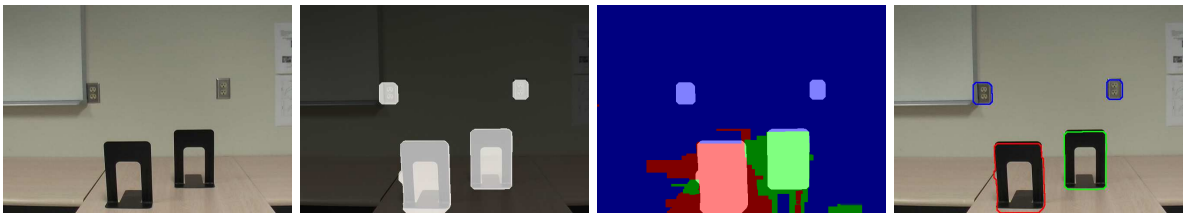
Machrouh, J. and Tarroux, P. (2005). Attentional mechanisms for interactive image exploration. *EURASIP Journal of Applied Signal Processing*, 14:2391–2396.

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 (to appear)*.

Noton, D. and Stark, L. (1971). Scanpaths in Eye Movements during Pattern Perception. *Science*, 171:308–311.

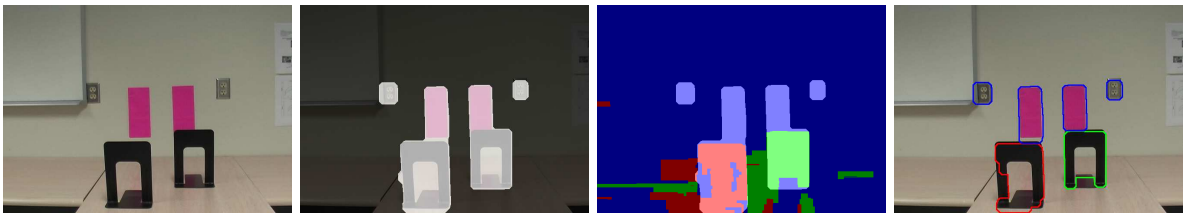
Stentiford, F. (2003). An attention based similarity measure with application to content-based information retrieval. In *Proceedings of the Storage and Retrieval for Media Databases Conference, SPIE Electronic Imaging*, Santa Clara, CA.

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



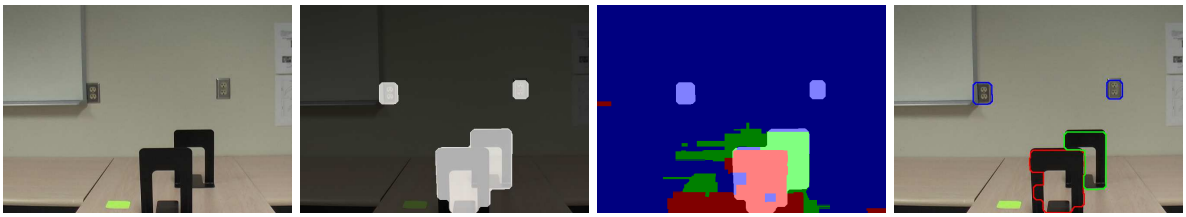
(a) (b) (c) (d)

Figure 3: Results for non-occluding salient objects in the foreground and with no distracting salient objects in the background: (a) original left stereo image, (b) highlighted ROI using saliency-based mask M, (c) saliency/depth-based ROI mask, and (d) final ROIs highlighted in the actual image.



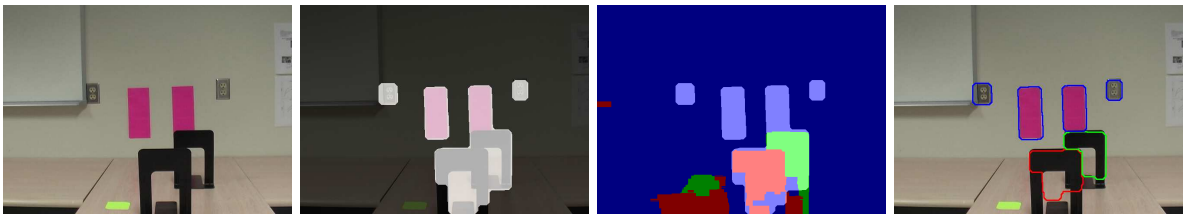
(a) (b) (c) (d)

Figure 4: Results for non-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) saliency/depth-based ROI mask, and (d) final ROIs highlighted in the actual image.



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

Figure 5: Results for occluding salient objects in the foreground and with no distracting salient objects in the background: (a) original left stereo image, (b) highlighted ROI using saliency-based mask M, (c) saliency/depth-based ROI mask, and (d) final ROIs highlighted in the actual image.



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

Figure 6: 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) saliency/depth-based ROI mask, and (d) final ROIs highlighted in the actual image.