A Novel Approach for Attention Points Extraction from Saliency Maps based on T-Junctions

Ekaterina Potapova, Michael Zillich, Markus Vincze
{potapova,zillich,vincze}@acin.tuwien.ac.at
Automation and Control Institute
Vienna University of Technology

Abstract

Object detection is of a vital importance in many computer vision and robotic tasks. Saliency maps can be used to extract attention points, that potentially indicate objects in the scene and therefore, can be further explored. For instance, attention points can serve as seed-points for object segmentation. However, it is important to guarantee the uniqueness and the quality of attention points. Therefore, the extraction of attention points is a challenging problem. In this paper, we propose a novel approach to extract attention points from saliency maps. Compared to several existing attention points extraction strategies, we show that the proposed strategy performs better in terms of the uniqueness of attention points and their proximity to the center of detected objects[1].

Keywords: *Visual Attention, Fixations, Saliency Maps.*

1. INTRODUCTION

Object detection is a crucial task in many applications of computer vision such as in robotics. One way to detect objects is to generate saliency maps and extract fixation points, so-called attention points, from them. These fixations can potentially identify or detect objects and can be used as seeds for segmentation [7] and further recognition.

One of the main challenges of attention points extraction mechanisms is inability to guarantee the uniqueness and the quality of them. Therefore, attention points extraction remains a challenging process. An obvious approach to evaluate the quality of attention points is to compare them to fixation maps built on human data. However, when the task at hand is to extract attention points suitable for a specific task (*e.g.* attention-driven segmentation), comparison to fixation maps is not very useful.

In this paper, we compare several existing attention points extraction strategies and evaluate them on different types of saliency maps in terms of the uniqueness and proximity of fixations to the center of the detected object. The later is important, since Vishwanath *et al.* [11] showed that line of sight lands near the center of gravity of the object. We also propose a novel approach to extract attention points from saliency maps based on T-Junctions. This attention strategy can be applied for saliency maps in which several disjoint connected components are available. We show that the proposed strategy performs better than existing strategies in terms of the above criteria.

The paper is structured as follows: In the next section, we describe algorithms for saliency map calculation, subsequently, attention points extraction strategies are discussed in detail. The following evaluation shows comparison of different strategies and in the end, we conclude our work with a discussion about future directions of research.

1The research leading to these results has received funding from the Austrian Science Fund (FWF) under grant agreement No. TRP 139-N23 InSitu and from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement No. 600623, STRANDS.

2. SALIENCY MAPS

We evaluated attention points extraction strategies on four different types of saliency maps: Attention based on Information Maximization (*AIM*) [1] (Fig.1,b), Graph-based visual saliency (*GBVS*) [3] (Fig.1,c), 2D Symmetry-based saliency (*SYM2D*) [5] (Fig.1,d), and 2.5D Symmetry-based saliency (*SYM2.5D*) [9] (Fig.1,e). In this section, we describe the basic principles of each saliency map.

2.1 Attention based on Information Maximization

Bruce *et al.* [1] proposed a model of bottom-up overt attention based on the principle of maximizing information sampled from the scene.

The model defines saliency by quantifying the self-information of each local image patch. Independent Component Analysis is performed on a large number of sampled patches to determine a suitable basis. The probability of observing a specific patch can be evaluated by independently considering the likelihood of each corresponding basis coefficient. Shannon's self-information measure, applied to the joint likelihood of statistics of the patch, provides an appropriate transformation between probability and the degree of information inherent in the local statistics. Therefore, saliency is determined as the self-information of each local image patch.

2.2 Graph-Based Visual Saliency

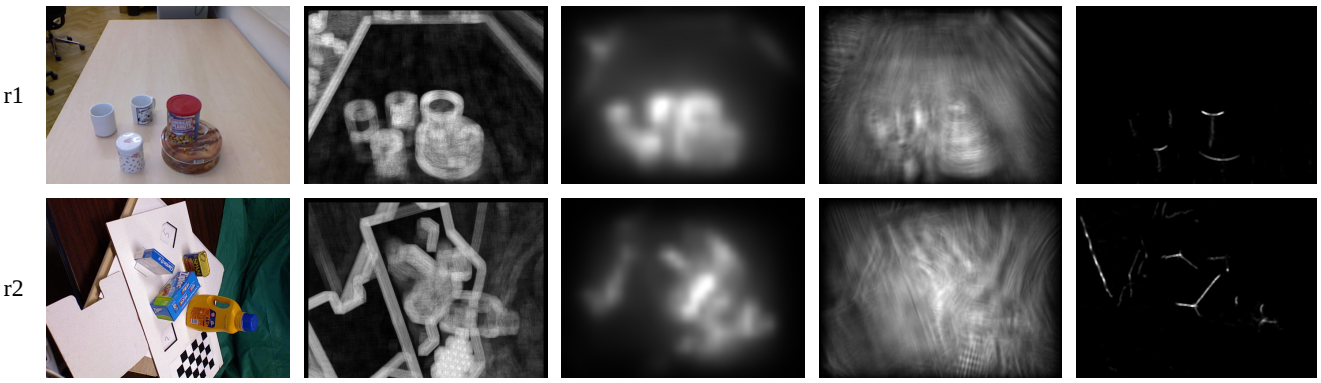
Harel *et al.* [3] described a simple and biologically plausible model for bottom-up saliency.

In the introduced model, activation maps are formed on certain feature channels (color, orientation), and then combined into the master saliency map. Activation maps are calculated using fully-connected directed graphs, where weights of the edges depend on the similarity between pixels they connect. Harel *et al.* [3] showed how to treat this graph as Markov chain. Therefore, activation map is an equilibrium state of the given Markov chain. Activation maps are normalized in a similar fashion using Markov chain, where the goal is to concentrate intensity of activation maps. Finally, normalized activation maps are combined into the master saliency map.

2.3 Saliency Map Based on 2D Symmetry

Kootstra *et al.* [5] proposed to use symmetry, one of the Gestalt principles for figure-ground segregation, to calculate saliency maps.

Saliency maps are built upon the local symmetry operator of Reisfeld *et al.* [10] and is extended to a multi-scale model similar to the contrast-saliency model [4]. A context free attentional



(a) Image *(b) AIM* *(c) GBVS* *(d) SYM2D* *(e) SYM2.5D*

Figure 1: Examples of images and saliency maps (Attention based on Information Maximization (*AIM*), Graph-based visual saliency (*GBVS*), 2D Symmetry-based saliency (*SYM2D*) and 2.5D Symmetry-based saliency (*SYM2.5D*)) for TOSD (r1) and WILLOW (r2) database.

operator, described by Reisfeld *et al.* [10], based on the intuitive notion of symmetry. The amount of local symmetry at a point is calculated as the sum of similarity measures between pixel pairs in the symmetric kernel centered at the point. Two points in the kernel form a pair if the middle point of the line connecting them is the center of the kernel. The similarity measure takes into consideration gradient directions and magnitudes, as well as distances between points.

2.4 Saliency Map Based on 2.5D Symmetry

Potapova *et al.* [9] extended principles introduced by Kootstra *et al.* [5] to be used in 2.5D space.

In this approach, saliency is calculated as the amount of reflective symmetry in a local patch. Reflective planes are determined as planes perpendicular to the directions of the principal components. A local patch is divided into two sub-patches by a reflective plane. The reflective symmetry measure includes such characteristics as the mean distance between two sub-patches, difference in mean depth values, and collinearity between mean normal vectors.

3. ATTENTION POINTS EXTRACTION STRATEGIES

In this section, we will describe different strategies employed to extract attention points from saliency maps. We are going to discuss the following strategies: Winner-Take-All (WTA [6]), Maximum Salient Region (*MSR* [2]), and T-Junctions.

3.1 Winner-Take-All

Lee *et al.* [6] showed that the Winner-Take-All (WTA) neural network can be used to simulate humans behavior of scene components prioritization while observing a scene.

Excitatory input neurons in the network are independent and received from a saliency map and each neuron excites its corresponding WTA neuron. All WTA neurons evolve independently of each other. The “winner" is the one that fires first (*i.e.* reaches threshold). This triggers complete reset of all WTA neurons. Attention points can be defined as points of location of the “winner" neuron. Firing of the “winner" neuron is followed by the shift of the Focus of Attention (FOA) and the local inhibition. FOA is shifted to be at the location of the “winner" neuron. Input neurons are inhibited at the new location of the FOA. Local inhibition prevents returning the FOA to the just attended location and allows the next most salient location to become the winner. This process is called ``Inhibition of Return" (IOR) and is described in [8]. All time constants, conductances, and firing thresholds used in the WTA model implemented in this paper are the same as in [4]. Attention points are extracted using down-sampled saliency maps.

3.2 Most Salient Region

Frintrop [2] proposed the detection of the Most Salient Region (*MSR*) for object detection. The point with the maximum saliency value (attention point) determines the most salience region.

Starting from the attention point (seed), the surrounding salient region is extracted by means of seeded region growing. *MSR* consists of all neighbors of the seed with saliency values that differ by at most 25% from the saliency value of the attention point. The focus of attention (FOA) is directed to the *MSR*. After that the *MSR* is inhibited and the next *MSR* is computed. Though this way of attention points detection is less biologically plausible, Frintrop [2] argues that equivalent results are achieved with fewer computational resources.

3.3 T-Junction Attention Points

We propose a new T-Junction (TJ) attention points extraction strategy. This strategy requires saliency maps in which each object is detected as an individual blob (*i.e. SYM2.5D*). Attention points are defined as junction points of multi-segment skeletons, or as mid-points in the case of single segments. Skeletons are calculated from connected components of saliency maps.

4. EXPERIMENTS

Comparison of different attention points extraction strategies applied to different types of saliency maps was done with respect to two metrics. The first metric is the *Hit Ratio* (*HR*), and the second metric is the *Distance to the Center* (*DC*). Experiments were performed on two databases: *Table Object Scene Database*[2] (*TOSD*, Fig.1,r1,a) and *Willow Garage Table Objects Database*[3] (*WILLOW*, Fig.1,r2,a). Objects in both databases were hand-labeled with polygon masks.

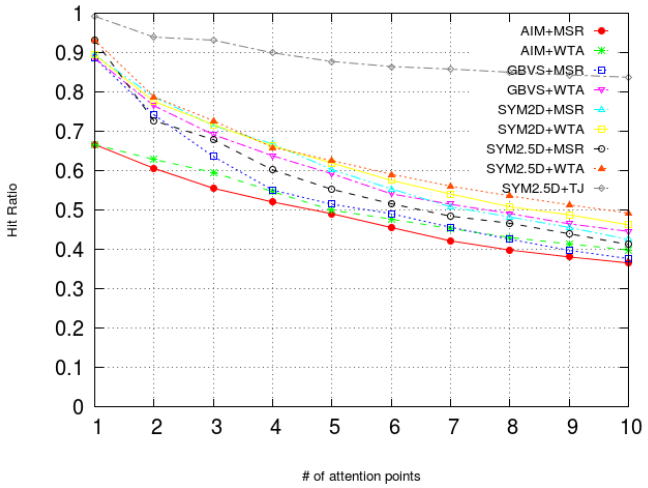
4.1 Hit Ratio

Hit Ratio (*HR*) shows how many different objects were covered by attention points with a given number of fixations, and is given by the percentage of unique attention points being situated inside different objects:

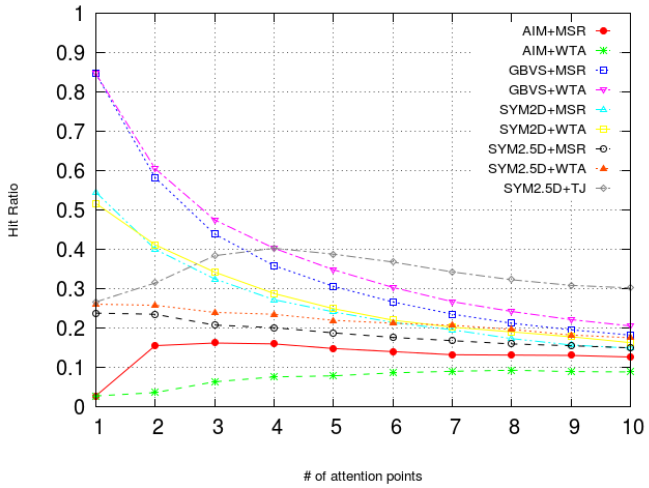
2 https://repo.acin.tuwien.ac.at/tmp/permanent/TOSD.zip
3 http://vault.willowgarage.com/wgdata1/vol1/solutions_in_perception/Willow_Final_Test_Set/

$$HR = n / N$$

where N is the number of fixations and n is the number of different attended objects. A perfect attention mechanism will hit every object exactly once, resulting in HR equaled to one.



(a) TOSD



(b) WILLOW

Figure 2: Hit Ratio (HR) against the number of extracted attention points for: (a) TOSD, (b) Willow. HR results are shown for different combinations of saliency maps and extraction strategies. As can be seen from the plot, the combination of SYM2.5D [9] with T-Junction extraction strategy results in the best performance.

We evaluate *HR* against the number of fixations. Fig.2,a and Fig.2,b show *HR* results averaged over all images in *TOSD* and *WILLOW* respectively. As can be seen from the plot, for *TOSD* 2.5D symmetry-based saliency maps together with attention points extraction strategy based on T-Junctions result in the increase of performance up to 60% in terms of the number of detected objects. The second best performance is achieved by extracting attention points from 2.5D symmetry-based saliency maps using Winner-Take-All extraction strategy. These results show that 2.5D symmetry-based saliency maps capture the main structure of table scenes with man made objects and direct attention to those objects.

For *WILLOW* database combination of 2.5D symmetry-based saliency maps and T-Junctions based extraction strategy result in low scores with *HR* for the first few attention points with improvement in performance up to 50% with a larger number of attention points. In *WILLOW* database only objects standing on the table were labeled, and neither tables, nor cardboards are considered as objects. On the other hand, 2.5D symmetry-based saliency operator gives a strong response for regions that are symmetrical in 3D space, for *e.g.* box and table corners. Combination of those two factors result in a high false detection rate.

Furthermore, we can conclude from plots in Fig.2, that *WTA* extraction strategy shows better performance, than *MSR* extraction strategy for a given saliency map.

4.2 Distance to the Center

The distance between the extracted attention point p and the center of the respective object c is defined as:

$$DC(p,c) = |p - c|_2$$

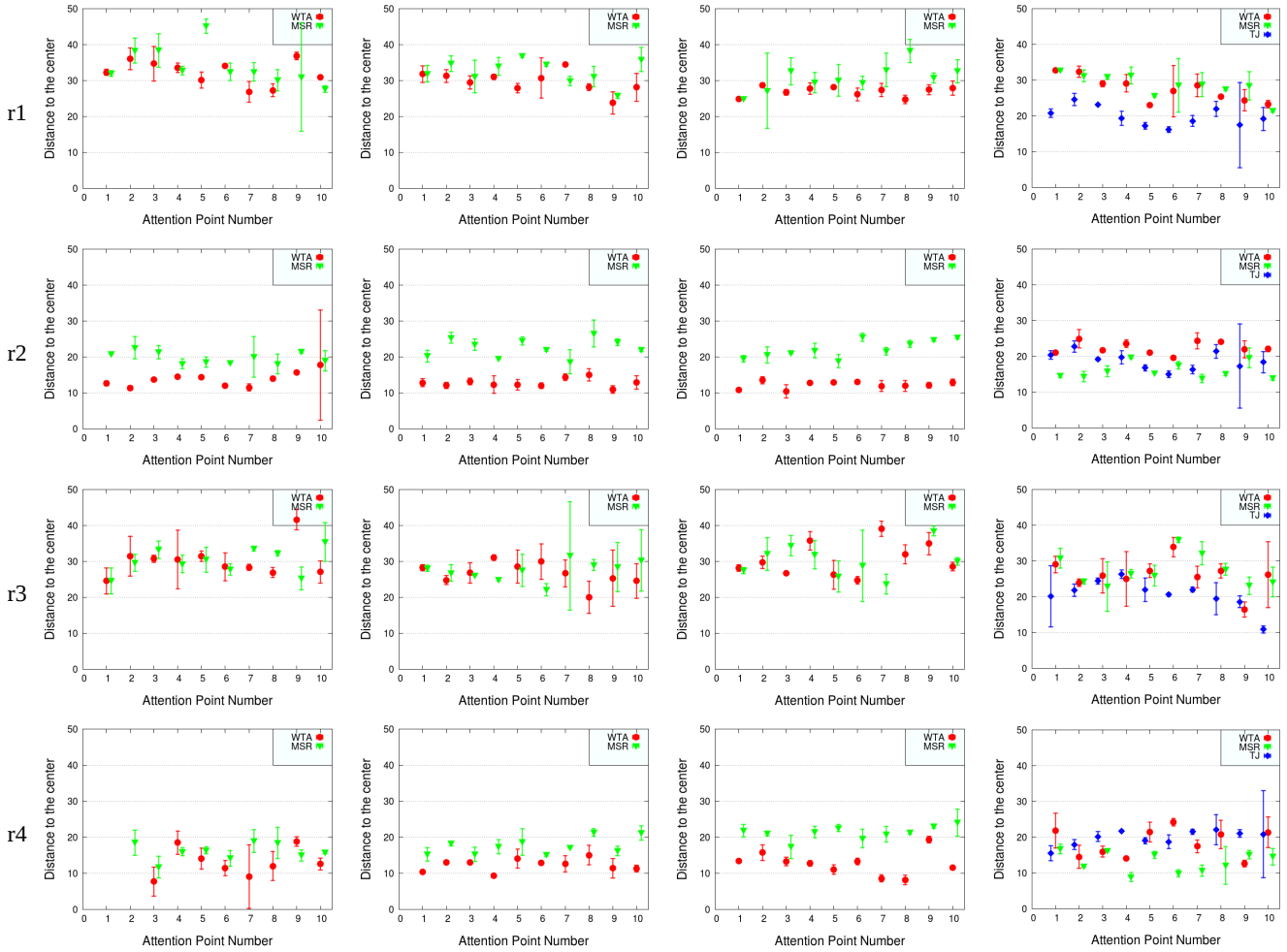
The centers of the respective objects represent physical centers of the visible parts of the objects. The *DC* shows the accuracy of attention points. The smaller the distance, the better is the detection quality of attention points.

Fig.3,r1-r2 and Fig.3,r3-r4 show evaluation of attention points with respect to the distance to object centers for *TOSD* and *WILLOW* respectively. Fig.3,r1 and Fig.3,r3 show evaluation results for attention points that were the first to detect objects. Fig.3,r2 and Fig.3,r4 show evaluation results for best attention points. The best detected attention point for an object is the one that is the closest to the center of the object.

A perfect attention mechanism would detect every object only once and directly at the center. As can be seen from plots in Fig.3 all detection strategies result in similar performance in terms of distance to the center for the first detected attention point. However, best attention points extracted by means of *WTA* strategy are situated closer to the center, than those detected by means of *MSR*. All extraction strategies show similar performance for 2.5D symmetry-based saliency map. This effect can be explained by the nature of 2.5D symmetry-based saliency in which salience blobs represent symmetry lines of objects. All described attention points extraction strategies detect attention points on these symmetry lines. Therefore, attention points are located close to each other.

5. CONCLUSION

In this paper we proposed a new strategy for extraction of attention points (*TJ*) and evaluated it against such strategies as Winner-Take-All (*WTA*) and Most Saliency Region (*MSR*) on several types of saliency maps (*AIM*, *GBVS*, *SYM2D*, *SYM2.5D*). Experiments show that the combination of 2.5D symmetry-based saliency map and T-Junctions extraction strategy performed up to 60% better than other combinations with respect to Hit Ratio criteria. Therefore, we can say that this combination can be specifically used to detect objects in cluttered table scenes containing man made objects. It was also shown that all extraction strategies give similar results with respect to the distance to the center of the object, when only first attention points are evaluated. However, the *WTA* strategy extracts attention points situated closer to the object center, when best attention points are evaluated. In future, we plan to modify the T-junction strategy to be applied not only for 2.5D symmetry-based saliency maps, but for other types of saliency maps as well.



(a) AIM *(b) GBVS* *(c) SYM2D* *(d) SYM2.5D*

Figure 3: Rows (r1) and (r3) show averaged distances from first attention points to the centers of detected objects for *TOSD* and *WILLOW* respectively. Rows (r2) and (r4) show averaged distances from best attention points to centers of the detected objects for *TOSD* and *WILLOW* respectively.

6. REFERENCES

[1] Bruce, N.D.B., Tsotsos, J.K. 2009. Saliency, Attention, and Visual Search: An Information Theoretic Approach, Journal of Vision 9, 3.

[2] Frintrop, S. 2006. VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, vol. 3899 of Lecture Notes in Computer Science. Springer.

[3] Harel, J., Koch, C., Perona, P. 2006. Graph-Based Visual Saliency. In NIPS, 545-552.

[4] Itti, L., Koch, C., and Niebur, E. 1998. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence 20, 11, 1254-1259.

[5] Kootstra, G., Bergstroem, N., and Kragic, D. 2010. Using Symmetry to Select Fixation Points for Segmentation. In 20th International Conference on Pattern Recognition (ICPR), 3894-3897.

[6] Lee, D. K., Itti , L., Koch, C., and Braun, J. 1999. Attention Activates Winner-Take-All Competition Among Visual Filters. Nature Neuroscience 2, 4, 375-381.

[7] Mishra, A. K., and Aloimonos, Y. 2009. Active Segmentation with Fixation. In IEEE 12th International Conference on Computer Vision (ICCV), 468-475.

[8] Posner, M. I., and Cohen, Y. 1984. Components of Visual Orienting. Attention and Performance X 32, 531-556.

[9] Potapova, E., Zillich, M., and Vincze, M. 2012. Local 3D Symmetry for Visual Saliency in 2.5D Point Clouds. In 11th Asian Conference on Computer Vision (ACCV), vol. 7724 of Lecture Notes in Computer Science. Springer, 434-445.

[10] Reisfeld, D., Wolfson, H., and Yeshurun, Y. 1995. Context Free Attentional Operators: the Generalized Symmetry Transform. International Journal of Computer Vision 14, 119-130.

[11] Vishwanath, D., and Kowler, E. 2004. Saccadic Localization in the Presence of Cues to Three-Dimensional Shape. Journal of Vision 4, 445-458.