Progressive Saliency-Oriented Object Localization Based on Interlaced Random Color Distance Maps

Maiko M. I. Lie, Hugo Vieira Neto, Humberto R. Gamba
Graduate Program in Electrical and Computer Engineering
Federal University of Technology – Paraná
Curitiba, Brazil
minian.lie@gmail.com, {hvieir, humberto}@utfpr.edu.br

Gustavo B. Borba
Department of Electronics
Federal University of Technology – Paraná
Curitiba, Brazil
gustavobborba@utfpr.edu.br

***Abstract*—The human visual system employs an information selection mechanism, visual attention, so that higher-level cognitive processes can be restricted to a potentially important subset of the incoming information. This mechanism is amenable to efficient computational implementation and, consequently, it has been incorporated into many technological applications. Among these applications is autonomous mobile robotics, in which efficient vision systems are paramount given their limited computational resources and energy autonomy requirements. Robots have employed visual attention for directing gaze and accelerating object detection, among other tasks. However, it has always been approached as a single rigid input-output stage. In this work, a bottom-up, unsupervised visual attention model based on progressive processing is presented. It adopts an incremental approach, in which a rough output is rapidly computed, and then successively refined, providing graceful-degradation, which might be particularly useful by robots based on the subsumption architecture. Progressive processing is achieved at a pixel saliency estimation level by adopting a method based on color distance to random pixel samples, and at a scale level through bidimensional interlacing. The proposed approach is assessed in the SIVAL dataset, and compared to other two visual attention models commonly employed in robot vision, presenting highly competitive performance.**

***Index Terms*—Visual attention, Saliency detection, Robot vision**

I. Introduction

Robot vision shares many of the issues that limit the human visual system. Among them is the inability to entirely process the incoming visual information in detail, due to limited processing capacity. From psychological experiments [1], [2], it is known that the human visual system copes with this limitation by employing a selective mechanism called *visual attention* that reduces the operation of posterior processing stages to a significantly smaller, potentially important, subset of the information. This allows rapid response to visual stimuli, despite the immense amount of information that is constantly presented to the visual system. Computational reproductions of this mechanism have been successfully applied in several vision-based technological applications such as image compression [3], and image retargeting [4]. The selective capabilities of visual attention are also applicable and especially important in robot vision, where rapid response is required, while delay associated with complex control systems and actuators might be involved. This is particularly true for autonomous mobile robots, where not only hardware capability is more limited, but parsimonious use of processing power is of vital importance for increased energy autonomy.

978-1-5386-0956-9/17/$31.00 © 2017 IEEE

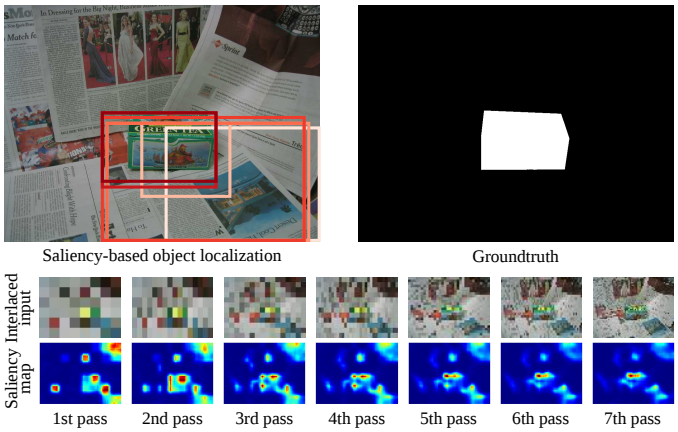

Fig. 1. Progressive object localization based on visual saliency. A rough saliency map is computed rapidly and then successively refined in a coarse-to-fine manner (bottom rows). Object localization using saliency maps computed from the proposed approach is indicated over the input image in red (top left). More refined passes are indicated by darker red. Each rectangle indicates the largest connected component of the thresholded saliency map.

Since information selection based on high-level, semantic, information is subjective and task-dependent, most research on computational visual attention has focused on its low-level aspects, namely *saliency detection*, which is the estimation of the distinctiveness of each location in the scene based on low-level image features (i.e. color, orientation, depth). Such a general, stimulus-based, approach is amenable to efficient computational implementation, an advantage that has made it widely employed for visual search space reduction prior to more time consuming tasks, such as object detection and recognition. Several computational methods have been proposed to estimate visual saliency, based on approaches such as graph-theory [5], Bayesian learning [6] and frequency-domain analysis [7]. From an input image, these methods compute a single *saliency map*, which is an image that indicates the degree of saliency in each scene location. The problem with most of these methods is usually twofold: (i) long execution time,

limiting their application as pre-processing filters, and (ii) rigid processing scheme, limiting their robustness in dynamic applications. While the former can be partially addressed by downsizing the input image, the latter has been mostly neglected, since most authors treat saliency detection as a single input-output stage, even if in practice they are comprised of several processing stages.

Psychological evidence suggests that visual attention efficiency is not constant, and might be significantly affected by processing capacity availability [8]. The consideration of this additional aspect – that the process of visual attention itself might be "truncated" due to limited time available for computation in a given circumstance – can lead to significant increase in the robustness of computational visual attention models. This suggests that it might be useful to model visual attention as a progressive process, capable of providing increasingly better results according to available time. While the usefulness of this aspect might vary for different applications, it is clearly advantageous for autonomous mobile robots, since they commonly operate on dynamic environments and require adaptive behavior to respond adequately in such conditions. More specifically, operating circumstances (e.g. higher priority behavior) might require subsumption (i.e. suppression, inhibition) [9] of vision-related behavior, making it highly desirable to have a visual attention system designed to provide useful output even if interrupted, since it might happen at any moment. Considering this, we propose a progressive visual attention model capable of fast output and successive refinement (Figure 1), and assess its performance for the task of object localization.

The proposed model computes a rough saliency map as soon as possible and successively improves it, in order to provide the most accurate output until a possible interruption. Unlike top-down, reinforcement learning methods [10], the proposed method provides progressive processing of bottom-up saliency estimation. It is based on a combination of saliency detection using random color distance maps [11] and interlacing based on the approach adopted for progressive display by the PNG (Portable Network Graphics) standard [12]. Results of object localization experiments on the SIVAL dataset show that the model compares favorably to other two widely used approaches that are commonly employed in robot vision. Through assessment in terms of precision, recall, F-measure and execution time, we show that the proposed approach is effective and can be significantly advantageous in resource-constrained systems, providing an additional level of robustness largely unexplored by previous computational visual attention methods.

II. RELATED WORK

A. Visual Attention in Robot Vision

An early application of visual attention in robot vision was presented by Scheier and Egner [13], in which a connectionist, intensity-based, bottom-up visual attention system was embedded in a mobile robot for object localization. In their experiments, large objects were detected and used as landmarks for navigation, which was successful on their simple experimental environment. This system was based on the computational visual attention model proposed by Goebel [14], which is an oscillatory neural model, that distinguishes different objects by means of their phase relation, i.e. an uncorrelated phase relation implied different objects.

Later applications on robot vision were mostly based on the highly successful model proposed by Itti, Koch and Niebur [15] – henceforth called IKN. This model was formulated around the *Feature Integration Theory* by Treisman and Gelade [1], in which visual attention is attracted to scene locations that are distinctive with respect to their surroundings in terms of low-level features, such as color, intensity and orientation.

Kismet, a robotic head designed for social human interaction [16], adopted a visual attention system based on the IKN model, extending it with motion saliency and face detection. Combined with the robot's motivational state, candidate locations indicated by the visual attention system were employed to guide its eye gaze. Vieira Neto and Nehmzow [17] employed IKN as an interest point detector in a novelty detection framework for an autonomous mobile robot, showing that this approach presents more consistent results than the multi-scale Harris corner detector [18]. Frintrop, Jensfelt and Christensen [19] employed an IKN-based visual attention model to detect and track scene locations in order to build landmarks in a visual SLAM application.

A more recent and very elegant visual saliency model was presented by Hou and Zhang [7], in which salient regions of the scene are computed as a difference in the frequency-domain. Based on natural image statistics literature, the authors found that a general image model might be estimated from a single image, more precisely, that the average-filtered log spectrum of an image is a reasonable approximation for the mean log spectra of several natural images. Thus, subtracting the log spectrum of the image from its average-filtered log spectrum results in the "innovative" content in the image, that is, information that is not common to natural images in general. This difference is called the *spectral residual*, and was shown to perform as an effective saliency map when transformed to the spatial domain, outperforming IKN in both accuracy and execution time [7]. Motivated by its short execution time, Rudinac and Jonker [20] employed the spectral residual in robot vision for the task of object localization. Besides the spectral residual of the intensity channel, as in its original formulation [7], the authors considered two additional color-opponent channels: red−green and blue−yellow, following the approach of Walther and Koch [21]. Maximally Stable Extremal Regions [22] were detected and clustered over the saliency map, indicating salient object locations.

B. The Psychological Plausibility of Varying Efficiency in Visual Attention

The ability of visual attention to ignore distractors have been associated to perceptual load by Lavie's *Load Theory of Attention* [8], which claims that the visual system exhausts its processing capacity on relevant information when it is overloaded, while processing capacity "spills over" distractors when it is underloaded. The original formulation of this theory proposed it as a resolution to the *early vs. late selection* debate,

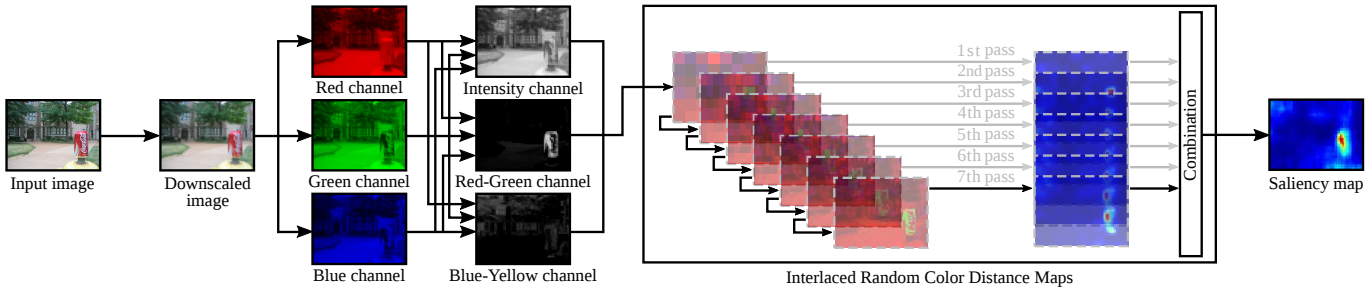

Fig. 2. The input image is uniformly downsized to a scale commonly assumed as appropriate for pre-attentive vision (64 pixels in width) [7]. From its RGB channels, an intensity and two color-opponent channels are computed. This representation is then progressively sampled through 7-pass bidimensional interlacing, which provides coarse-to-fine versions of the image. For a given interlacing pass, an intermediate saliency map estimated from color distances to random samples of the image is computed before sampling the next pass. Sampling from both interlacing and color distance computation can be interrupted, in which case the data processed until then is combined into a saliency map.

which questioned whether distractors were filtered based on semantic information or prior to any semantic interpretation. Considering that both of these seemingly opposing views are supported by experimental evidence, the Load Theory of Attention proposed a hybrid framework based on processing capacity: selection is early or late according to how much perceptual processing capacity is available – early if overloaded, late otherwise. More recent work has argued that this theory does not address late selection conditions at all, as Benoni and Tsal [23] argued that what it does is to provide evidence for the varying efficiency of *early* attention. In this work, we are not concerned about the precise interpretation of the varying efficiency of visual attention – only that it is very akin to what happens on the human visual system and that this principle might be leveraged to design more robust computational visual attention systems. For a detailed discussion of the current research regarding Lavie's Load Theory of Attention and its competing interpretations, the reader is referred to [23]–[26].

III. PROPOSED METHOD

The proposed method (Figure 2) adopts a progressive processing approach with two main components: (i) saliency estimation using *random color distance maps* and (ii) coarse-to-fine decomposition using *bidimensional interlacing*. The former provides control of processing at pixel saliency-estimation, while the latter provides control at a scale level.

A. Saliency Estimation

The *random color distance map* approach is a fast method based on color distances to random samples [11]. This method allows progressive processing at a saliency estimation level. More precisely, given an input image I, the saliency $S(x, y)$ of each pixel $I(x, y)$ is estimated as:

$$S\,(x,y) = \sum_{\forall\,(x_r,\,y_r)\,\in\,I_R} || \,I(x,y)\;-\;I(x_r,y_r)\,||, \qquad (1)$$

where I_R is a set of pixel locations randomly sampled from I. The random set I_R is resampled for each $S(x, y)$ – it is through this sample that progressive processing is achieved. All pixels $I(x, y)$ have their color distance computed to a single pixel of their instance of I_R, then to the next pixel, and so on until the set size N_R. In this manner, accuracy is improved more homogeneously throughout the image and all pixels have a saliency estimate as soon as possible.

1	6	4	6	2	6	4	6
7	7	7	7	7	7	7	7
5	6	5	6	5	6	5	6
7	7	7	7	7	7	7	7
3	6	4	6	3	6	4	6
7	7	7	7	7	7	7	7
5	6	5	6	5	6	5	6
7	7	7	7	7	7	7	7

Fig. 3. Sampling pattern for 7-pass (Adam7) interlacing. The numbers indicate in which pass the pixel each position will be sampled. If interlacing is interrupted, missing values are obtained by nearest neighbor interpolation.

For salient object detection, which attempts to obtain segmentation-level accuracy, the random color distance map was originally joint upsampled [11]. This was done to address the noisy output resulting from the adoption of small values of N_R, which in turn was done to speed up execution. Since we are concerned with object localization, which does not require segmentation-level accuracy, a Gaussian low-pass filter (5×5 support) is adopted instead, further improving execution time. Additionally, unlike the original formulation by Lie and colleagues [11] that operated in the CIELAB color space, computation is done in an intensity channel and two color-opponent channels (i.e. red−green and blue−yellow) following the more computationally efficient approach by Rudinac and Jonker [20].

B. Bidimensional Interlacing

Progressive processing at scale level is provided by bidimensional interlacing, for which the *7-pass (Adam7) interlacing* pattern [12] was chosen. Adam7 is the approach adopted for progressive display in the PNG (Portable Network Graphics) format. This interlacing scheme is defined by a 8×8 sampling pattern (Figure 3), which is repeated through the entire image and performed in seven passes.

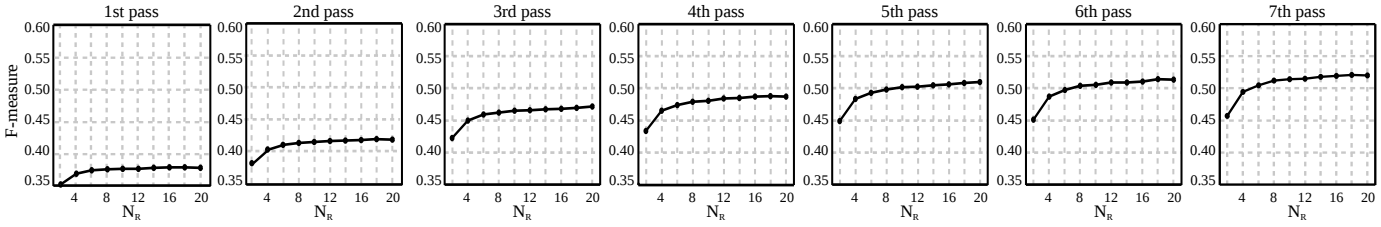

Fig. 4. Effect of the size N_R of the random sample in saliency estimation. Despite larger values marginally increasing accuracy in later passes, there is little improvement for $N_R \geq 8$. This is a convenient characteristic, since the execution time of saliency estimation increases linearly with respect to N_R.

The result of bidimensional interlacing using the 7-pass approach is roughly equivalent to a multi-scale decomposition. However, unlike other multi-scale decompositions such as the popular Gaussian pyramid [27], Adam7 interlacing proceeds directly in a coarse-to-fine manner, whereas a Gaussian pyramid must be entirely built in a fine-to-coarse manner before coarse scales can be used. For progressive processing, coarse-to-fine is preferred since its purpose is to compute a rough output as soon as possible, so that a useful output is available even if posterior refinement steps must be interrupted. Moreover, a Gaussian pyramid is built from successive filtering, while interlacing is computed simply as a sampling pattern, being more computationally efficient since it does not require any filtering at all. As a downside, unlike a Gaussian-scale space, interlacing does not follow the scale-space axioms nor is optimal in any mathematical sense. Since this limitation does not compromise the output, the proposed method adopts the interlacing approach for its implementation simplicity and computational efficiency.

C. Combination

The saliency map at a given instant is computed as the weighted combination of the saliency maps for all passes computed until then. The weights are defined so that earlier passes contribute less to the final result, since they were computed from significantly less data than later passes and consequently present more uncertainty. The combined saliency $S_c(x, y)$ of each pixel $I(x, y)$ is defined by Equation 2:

$$S_c(x, y) = \frac{1}{N} \sum_{n=1}^{N} n \cdot S_n(x, y), \tag{2}$$

where S_n is the saliency map for the nth pass of the interlacing, and N indicates the last pass computed at the moment of combination. While more elaborate approaches are possible, a simple approach is desirable, since the proposed method is meant for dynamic applications which require fast output.

IV. EXPERIMENTS

The proposed method was assessed in terms of precision, recall, F-measure ($\beta^2 = 0.3$) [28] and execution time, for the task of object localization. The experiments were executed in an Intel Core i7-860 2.80 *GHz* CPU with 4 *GB* RAM, using the publicly available SIVAL (Spatially Independent, Variable Area and Lightning) dataset [29]. It contains 24 object categories with 60 images each, totaling 1500 images with 1024×768 pixels in size, presenting changes in background, perspective, scale, and lightning conditions. Because this dataset does not originally provide ground-truth images, the manually extracted object binary masks used in the work of Borba and colleagues [30] were adopted as ground-truth.

TABLE I
DATA SAMPLED (%), ACCURACY AND EXECUTION TIME FOR EACH PASS.

Pass	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Data sampled	2%	4%	7%	13%	25%	50%	100%
F-measure	0.38	0.41	0.44	0.45	0.47	0.48	0.49
Exec. time (s)	0.02	0.03	0.04	0.06	0.09	0.11	0.14

Unlike some popular datasets used in visual attention assessment [28], [31], the SIVAL dataset is more suitable for the context of robot vision since it does not present center-bias (i.e. tendency for the objects to be in the center of the image). In images such as photographs, it is reasonable to assume that the object of interest is roughly centered due to photography composition principles such as the rule-of-thirds [32]. In robot vision this is not the case, since objects might be in the boundaries of the image due to robot movement.

A. Size of the Random Sample

Saliency estimation using random color distance maps is advantageous due to its computational complexity. It has linear complexity, with a constant factor proportional to the size N_R of the random samples used for distance computation [11]. To determine an adequate value for N_R, its F-measure was computed for several values for each pass, as shown in Figure 4. The results show that most of the accuracy is achieved for a small sample size, with little improvement after approximately $N_R = 8$. Later passes benefit more from a larger set size, but only marginally. Considering this, $N_R = 8$ was adopted for all passes.

B. Efficacy of Progressive Processing

One of the main advantages of the proposed method is that it is capable of computing a rough saliency map as soon as possible. In this manner, it has a guaranteed useful output even if it is truncated prior to significant refinement. However, this assumes that the initial saliency map is accurate enough to be useful. To verify whether this is the case, the accuracy of the output at the end of each pass was assessed, as shown in Table I. Considering the best accuracy possible with the

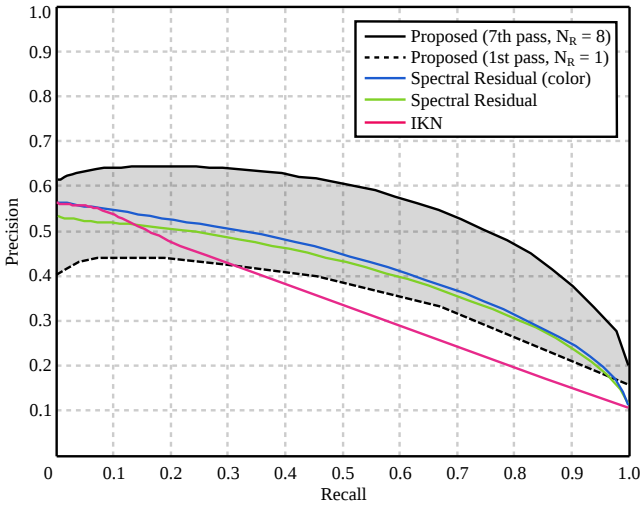

Fig. 5. Precision vs. Recall curves for the compared methods. The proposed method presents competitive accuracy even in the first pass, significantly surpassing all compared methods by its last pass. The shaded region indicates the accuracy range of the intermediate passes of the proposed method.

proposed method (i.e. N_R distances per pixel on all passes), it seems that the assumption is reasonable, since a very significant amount of the final accuracy of the method (i.e. F-measure 0.38 of 0.49) is achieved during the first pass. Moreover, the first pass achieves this while sampling only 2% of the image.

These results show that the accuracy of the initial pass is substantial compared to that of all passes. However, this does not guarantee that the method is advantageous for object localization. What remains is to show that the range of accuracy achievable by the method is competitive other state-of-the-art methods employed for this task. This is shown in the comparative analysis that follows.

C. Comparative Analysis

The proposed method is assessed and compared to other two approaches, IKN [15] and spectral residual (SR) [7]. This comparison is not meant to be exhaustive – the choice of algorithms was motivated by their prevalence as basis for visual attention systems on robot vision applications [16]– [20]. In the case of the spectral residual, it was considered both in its original intensity-based formulation and in the color derivation proposed by Rudinac and Jonker [20]. A qualitative comparison is presented in Figure 6.

The precision vs. recall curves for quantitative comparison are presented in Figure 5. The shaded region indicates the range of curves where the intermediate saliency maps of the progressive approach occur. As can be seen, during the first pass the proposed method is already competitive with the compared methods, while in its last pass it significantly surpasses all of them. In terms of execution time (Table II), the first pass of the proposed method has a performance equivalent to the intensity-based spectral residual, despite being less accurate. The color-based spectral residual is slightly slower, since it computes two additional color-opponent channels with respect to its intensity-based counterpart. Since it provides a subtle

TABLE II
ACCURACY AND EXECUTION TIME OF THE COMPARED METHODS.

	Proposed		SR	SR (color)	IKN
	1st pass	7th pass			
F-measure	0.38	0.49	0.45	0.46	0.33
Exec. time (s)	0.02	0.14	0.02	0.03	0.42

increase in accuracy, the additional execution time might be small enough so that it is still advantageous. The last pass of the proposed method, corresponding to the most accurate of the compared methods, has an average execution time of 0.14 seconds, which is significantly slower than the spectral residual approach. However, this is still within the time frame expected of bottom-up visual attention, which has been reported to take approximately 0.15 seconds in the human visual system [2]. IKN is the least accurate and also the most time-consuming of the compared methods. This does not mean that it is generally inefficient, but suggests that it might be more adequate for tasks such as gaze prediction than object localization.

V. CONCLUSIONS

While progressive processing has been largely unexplored in visual attention research, it provides very significant advantages for autonomous mobile robots. This is more explicit in systems that employ a subsumption architecture [9], where complex behavior occurs as consequence of the interaction of simpler ones that occur in parallel. In such an architecture, behaviors must be designed under the premise that they may be subsumed by higher priority behavior. In this case, the capability of providing a useful output as soon as possible through progressive processing, so that any interruptions only affect its refinement, is highly desirable. Thus, this paper presented such an approach for bottom-up visual attention and assessed it for the task of object localization.

The proposed approach has shown to be effective in terms of precision, recall, F-measure and execution time in the SIVAL dataset. Moreover, it demonstrated competitive accuracy and computational performance with respect to models commonly employed in robot vision, providing an output as soon as the fastest method and being capable of refinement until it is more accurate than all of them – within the time reported to be taken by the human visual system [2].

The method presented is a proof-of-concept that, despite being successful, can be significantly improved. In future work, we intend to generalize the interlacing approach and assess the method for any number of passes. For a proof-of-concept, the seven passes adopted by the PNG standard was reasonable, but it does not provide the granularity necessary for a more detailed analysis of the effect of different scales and amount of sampled data. We also intend to employ the proposed approach in an object detection and recognition framework to assess the effectiveness of its visual search space reduction. Some authors have achieved very promising results using the spectral residual recently [33], encouraging further investigation of the capabilities of our approach.

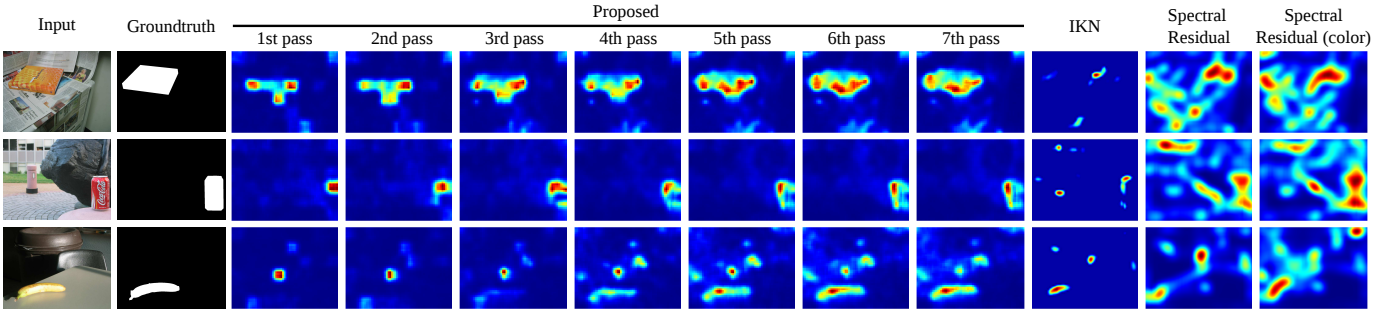

Fig. 6. Examples of saliency maps computed from the compared methods. IKN outputs mostly sparse small regions, which explains its low accuracy. The spectral residual is less sparse, but highlights several unimportant regions. The proposed approach produces increasingly accurate output, without significant increase in false positives.

ACKNOWLEDGMENT

The authors acknowledge the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) for the financial support of this work.

REFERENCES

[1] A. M. Treisman and G. Gelade, "A Feature-integration Theory of Attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97 – 136, 1980.

[2] J. Theeuwes, "Top-down and Bottom-up Control of Visual Selection," *Acta Psychologica*, vol. 135, no. 2, pp. 77 – 99, 2010.

[3] N. Ouerhani, J. Bracamonte, H. Hügli, M. Ansorge, and F. Pellandini, "Adaptive Color Image Compression Based on Visual Attention," in *Proc. of the International Conference on Image Analysis and Processing*, 2001, pp. 416–421.

[4] J. Sun and H. Ling, "Scale and Object Aware Image Retargeting for Thumbnail Browsing," in *Proc. of the IEEE International Conference on Computer Vision*, Nov 2011, pp. 1511–1518.

[5] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proc. of the International Conference on Neural Information Processing Systems*. MIT Press, 2006, pp. 545–552.

[6] Y. Xie and H. Lu, "Visual Saliency Detection Based on Bayesian Model," in *Proc. of the IEEE International Conference on Image Processing*, 2011, pp. 645–648.

[7] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[8] N. Lavie, "Perceptual Load as a Necessary Condition for Selective Attention," *Journal of Experimental Psychology: Human perception and performance*, vol. 21, no. 3, p. 451, 1995.

[9] R. Brooks, "A Robust Layered Control System for a Mobile Robot," *IEEE Journal of Robotics and Autom.*, vol. 2, no. 1, pp. 14–23, 1986.

[10] M. T. López, M. A. Fernández, A. Fernández-Caballero, J. Mira, and A. E. Delgado, "Dynamic Visual Attention Model in Image Sequences," *Image and Vision Computing*, vol. 25, no. 5, pp. 597 – 613, 2007.

[11] M. M. I. Lie, G. B. Borba, H. Vieira Neto, and H. R. Gamba, "Fast Saliency Detection Using Sparse Random Color Samples and Joint Upsampling," in *Proc. of the Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 217–224.

[12] T. Boutell and et al., "PNG (Portable Network Graphics) Specification," IETF, RFC 2083, 1997. [Online]. Available: https://tools.ietf.org/html/rfc2083

[13] C. Scheier and S. Egner, "Visual Attention in a Mobile Robot," in *Proc. of the IEEE International Symposium on Industrial Electronics*, vol. 1, 1997, pp. SS48–SS52.

[14] R. Goebel, "Perceiving Complex Visual Scenes: an Oscillator Neural Network Model that Integrates Selective Attention, Perceptual Organisation, and Invariant Recognition," *Advances in Neural Information Processing Systems*, pp. 903–903, 1993.

[15] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] C. Breazeal and B. Scassellati, "A Context-dependent Attention System for a Social Robot," in *Proc. of the International Joint Conference on Artificial Intelligence*, 1999, pp. 1146–1151.

[17] H. Vieira Neto and U. Nehmzow, "Visual Novelty Detection with Automatic Scale Selection," *Robotics and Autonomous Systems*, vol. 55, no. 9, pp. 693 – 701, 2007.

[18] K. Mikolajczyk and C. Schmid, "Indexing Based on Scale Invariant Interest Points," in *Proc. the IEEE International Conference on Computer Vision*, vol. 1, 2001, pp. 525–531.

[19] S. Frintrop, P. Jensfelt, and H. Christensen, *Simultaneous Robot Localization and Mapping Based on a Visual Attention System*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 417–430.

[20] M. Rudinac and P. P. Jonker, "Saliency Detection and Object Localization in Indoor Environments," in *Proc. of the International Conference on Pattern Recognition*, 2010, pp. 404–407.

[21] D. Walther and C. Koch, "Modeling Attention to Salient Proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395 – 1407, 2006.

[22] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide-baseline Stereo from Maximally Stable Extremal Regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004.

[23] H. Benoni and Y. Tsal, "Conceptual and Methodological Concerns in the Theory of Perceptual Load," *Frontiers in Psychology*, vol. 4, p. 522, 2013.

[24] Y. Tsal and H. Benoni, "Diluting the Burden of Load: Perceptual Load Effects are Simply Dilution Effects," *Journal of Experimental Psychology: Human Perc. and Performance*, vol. 36, no. 6, p. 1645, 2010.

[25] N. Lavie and A. Torralbo, "Dilution: a Theoretical Burden or Just Load? A Reply to Tsal and Benoni (2010)," *Journal of Experimental Psychology: Human Perc. and Performance*, vol. 36, no. 6, pp. 1657–64, 2010.

[26] Y. Tsal and H. Benoni, "Much Dilution Little Load in Lavie and Torralbo's (2010) Response: A Reply," *Journal of Experimental Psychology: Human Perc. and Performance*, vol. 36, no. 6, pp. 1665–1668, 2010.

[27] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid Methods in Image Processing," *RCA engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[28] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned Salient Region Detection," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597 – 1604.

[29] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts, "Localized Content Based Image Retrieval," in *Proc. of the ACM SIGMM International Workshop on Multimedia Information Retrieval*. ACM, 2005, pp. 227–236.

[30] G. B. Borba, H. R. Gamba, O. Marques, and L. M. Mayron, "Extraction of Salient Regions of Interest Using Visual Attention Models," *Proc. of the SPIE – IS&T Electronic Imaging*, vol. 7255, pp. 1–12, 2009.

[31] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global Contrast Based Salient Region Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[32] L. Mai, H. Le, Y. Niu, and F. Liu, "Rule of Thirds Detection from Photograph," in *Proc. of the IEEE International Symposium on Multimedia*, 2011, pp. 91–96.

[33] G. Silva, L. Schnitman, and L. Oliveira, "Constraining Image Object Search by Multi-scale Spectral Residue Analysis," *Pattern Recognition Letters*, vol. 39, pp. 31 – 38, 2014.