Visual Search for Object Instances Guided by Visual Attention Algorithms

Rafael Galvão de Mesquita and Carlos Alexandre Barros de Mello
Centro de Informática - Universidade Federal de Pernambuco
Recife, PE, Brasil
{rgm,cabm}@cin.ufpe.br

***Abstract*—In this Ph.D. research[1], we have proposed three visual search algorithms that connects visual attention and object recognition. The first, inspired by the bottom-up mechanism of visual attention, uses the saliency of the analysed scene to guide the search, prioritizing locations that are more salient; in order to do this, a saliency detector was developed. In the second method, based on the top-down mechanism of visual attention, the visual search is driven by characteristics of the searched object. In the third algorithm, the bottom-up and top-down visual searches are combined to improve the results of both approaches. Moreover, a modified version of SURF (Speeded-Up Robust Features) recognition algorithm was introduced so that the recognition occurs iteratively in the scene. Quantitative results showed that our saliency detector outperformed other nine algorithms and that our visual search proposal reduces the recognition time to 44% of the time achieved by SURF without our method.**

I. Introduction

Although the human brain receives, as signals, a huge amount of visual stimuli at every moment, human beigns are able to visually interact with the environment efficiently for most of our daily tasks. In addition to the problem of treating all the amount of input signals received by the brain, the task of recognizing objects and understanding scenes becomes even more difficult if one considers that each component of a given visual stimuli received from the environment needs to be compared to a large variety of known signals represented in memory [1].

Among the characteristics of the Human Visual System (HVS) responsible for its efficient interaction with the environment, we highlight the attentive mechanism, which is responsible for focusing the analysis of a scene in locations that are, somehow, selected as important. Thereby, thanks to the visual attention, the HVS is able to focus the analysis in a selected location, instead of processing an entire scene at once. Then, after processing the attended region, another location may be selected by the attentive mechanism, and this process repeats iteratively until a stop criterion is reached.

Visual attention can be driven in a bottom-up or in a top-down way. In the first one, the attended regions are defined unconsciously, based only on low level characteristics of the scene, as intensity, orientation or colour [2]. If, for a given location, one of these characteristics distinguishes from the rest of the scene or, at least, from its local neighbourhood, this region probably has a high *saliency* value and it tends to be prioritized during the analysis of the scene. On the other hand, a location that does not distinguish from its neighbourhood has a low saliency value and receives a low prioritization in the analysis of the scene.

Alternatively, according to the top-down mechanism of visual attention, cognitive factors, like knowledge or the objective of locating a specific object, have influence on the definition of which locations of the scene are focused. For example, car drivers are more likely to note petrol stations than pedestrians; or if someone is looking for an object of a specific colour, it is expected that regions with similar colour attracts his attention more than other locations [3].

Similarly to the human brain, a computer vision algorithm may also prioritize some regions of a scene to achieve a more efficient visual interaction with the world. This idea is especially important for complex scenarios, when certain conditions can not be assumed, like the main colours of the background or the positions of foreground elements.

The main goal of this Ph.D Thesis [4] is the development of a visual search model inspired by the visual attention mechanism of the human visual system. The proposal is experimented in the task of detecting the presence of an object in a scene, with the expectation of decreasing its recognition time. In order to achieve its objective, this research proposed: (i) a visual search model that can be guided by the bottom-up or by the top-down mechanisms, or by both; (ii) a modified version of the classic SURF recognition algorithm, that makes possible that the recognition of an object occurs iteratively and separately for each scene location selected by the visual attention mechanism; (iii) an algorithm to measure the saliency value of the pixels in an image, in order to guide the bottom-up search mode; and (iv) an algorithm that prioritizes regions of a scene according to how similar their characteristics are in relation to the characteristics of the searched object, in order to guide the top-down visual search.

The rest of this paper is organized as follows. Section 2 reviews related works. In Section 3, the scientific contributions of our research are presented along with their experimental results. Finally, Section 4 concludes the paper.

[1]Ph.D. thesis conducted between March 2013 and February 2017

II. Related Works

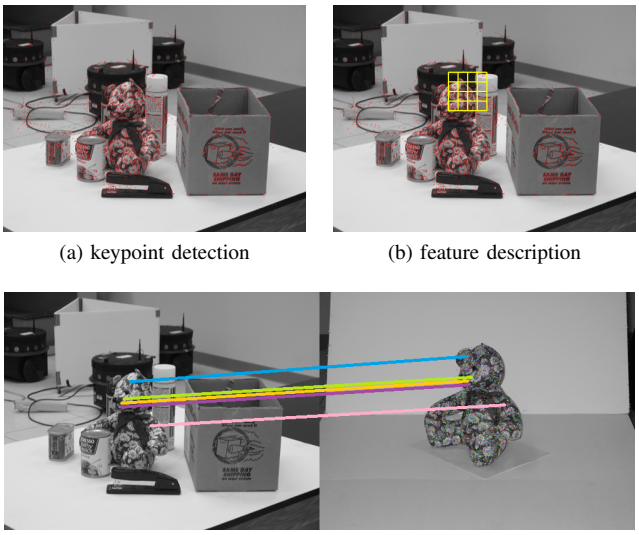
A. Recognition of Object Instances

Our Thesis applied new visual search methods in the task of recognizing object instances. This is a subtype of object recognition in which it is aimed to recognize rigid instances of an object. Although the generalization power of the recognition is not explored in this kind of problem, this is an important preprocessing step frequently used for higher level applications, as class recognition [5], landmark recognition [6], content based image/video retrieval [7] [8] [9], and others. Instance recognition algorithms usually deal with challenges such as changes in rotation, 3D viewpoint, scale and illumination.

The recognition of object instances can be decomposed into three main steps: keypoint detection, feature description and matching. The first one has the goal of detecting locations that are easily identified in other images containing the same object in different situations. In order to do this, blob-like structures (bright regions that distinguishes from a darker background, or the opposite) are detected in the image analysed. This step usually considers a representation of the image into multiple scales (*scale-space function*) followed by extrema point detection in the scale-space function. Due to this treatment, the recognition tends to become invariant to changes in scale; on the other hand, this procedure makes this step computationally very expensive, as it is needed to deal with multiple scale representations for the same image.

In the feature description step, the characteristics of the neighbourhood of each keypoint are extracted and a descriptor vector is built. As an object may appear in different orientations, and may be affected by changes in illumination, a descriptor should be, at the same time, invariant to these transformations and discriminative in relation to other objects. This stage requires the estimation of a dominant orientation to achieve orientation invariance, what also contributes to increase the recognition processing time. Thus, the descriptor is then built according to the estimated orientation using the features extracted from the keypoint's vicinity.

After the construction of the descriptors of all keypoints, a matching can be made between the descriptors extracted from the scene and the descriptors stored in a database representing the searched object. Figure 1 illustrates the three steps of object recognition described. After the matching step, one can also verify if the matched keypoints from both images are geometrically consistent (RANSAC [10] can be used for this).



(a) keypoint detection (b) feature description

(c) matching

Fig. 1. Object recognition main steps. In (a) and (b), red pixels highlight the detected keypoints. In (b), the yellow grid represents the neighbourhood of a keypoint used to build its descriptor vector. In (c), the descriptors from both images are matched.

B. Saliency detection

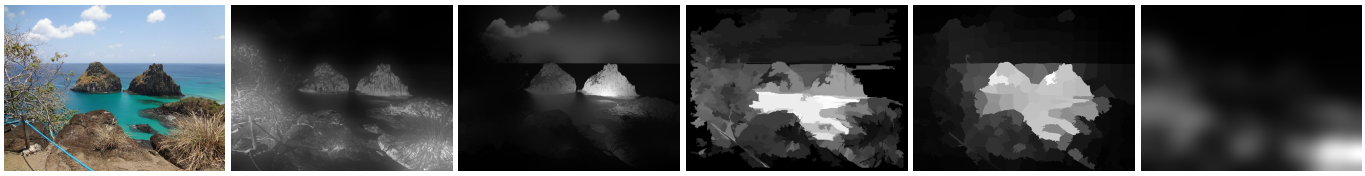
The saliency of the pixels or regions of an image is computed by saliency detection algorithms. These methods produce, as a result, a grayscale image, refereed as a *saliency map*; this is usually done without previous information about the analysed image, as its background or the position of its objects. In a saliency map, light gray tones represent higher saliency values, while darker tones indicate that a pixel has a lower saliency. Figure 2 presents examples of saliency maps produced by the saliency detector proposed in our Thesis and by other different approaches.

One can find in the literature saliency detection algorithms based on different concepts. For example, the method proposed by Itti, Koch and Niebur [16] aims to be biologically plausible. Others are based on colour differences, such as the method proposed by Achanta *et al.* [17], that computes the colour distance between the main background of the scene and its salient objects, or such as the proposal of Cheng *et al.* [13], that initially segment the regions of the image and then computes saliency based on the colour differences and on the spatial distance between the segmented locations.

The works of Perazzi *et al.* [12] and Cheng *et al.* [18] also segment the input image, but they combine the concepts of unity and spatial distributions of the segmented regions. There are also methods that compute saliency based on the Discrete Cosine Transform (Hou, Harel and Koch [15]), and on an image representation based on a Markov Chain (Jiang [14] *et al.*). One can find a review and comparison of saliency methods in the works of Borji *et al.* [19] [20].

C. Top-down visual attention methods

Unlike saliency detection, fewer algorithms based on the top-down mechanism of visual attention have been proposed in the literature, since top-down techniques are usually less expansible, more complex to design and also more time consuming [21], [22]. Other common problem of top-down methods is that they frequently direct the search towards some characteristics assuming that they attract human attention [23]. For example, person, car, and face detectors algorithms are used to train top-down attention algorithms in the methods proposed by Liu [2] and Borji [21]. This kind of approach is limited to situations in which good performance detectors of the searched object already exist, as it is the case of



(a) Original image (b) Mesquita [11] (c) Perazzi [12] (d) Cheng [13] (e) Jiang [14] (f) Hou [15]

Fig. 2. Example of an input scene (a), saliency maps produced by our method (b) and by algorithms proposed by other authors (c-f).

face detectors. Moreover, directing the search towards objects assumed as attractive may increase the processing time of the search if the object used to train the top-down attention algorithm does not correspond to the searched object.

Lee *et al.* [24] proposed a unified visual attention model (named as UVAM), that uses both top-down and bottom-up information to speed-up object recognition. Bottom-up attention is computed using the model proposed by Itti, Koch and Niebur [16] and the top-down mechanism is evaluated using partial results of the steps of keypoint detection and description of object recognition. This algorithm contains a preliminary recognition step, in which regions are prioritized based on high level information. Then, the saliency map is combined with the result of the preliminary recognition so that a region of interest is selected as the attended region, where a detailed recognition occurs. For the next iterations of the search, the results of the (i) saliency map, (ii) preliminary recognition and (iii) detailed recognition are combined to define the new attended region where the complete recognition occurs.

III. Scientific Contributions

The visual search proposed in this PhD Thesis differentiates from other proposal in some aspects, which are presented in this section. First, we propose a saliency detector algorithm that achieved superior results than state-of-the-art methods in the task of guiding the visual search. Very frequently, the method proposed by Itti *et al.*, although computationally expensive, is used for this task. The second main difference of our proposal is the fact that recognition occurs separately and iteratively for each focused region of the scene, thanks to the proposal of a patch-based version of SURF recognition algorithm. The third difference to be highlighted is that our work also presented a top-down mechanism that, differently from other techniques: (i) does not require any specific detector for the searched object, (ii) is efficient in terms of processing time, since most of its computation time is executed as part of the recognition process, and (iii) it also does not assume that certain object, as faces or persons, are attractive.

A. Patch-based SURF

Differently from classical recognition algorithms, like SURF [25] or SIFT [26], the patch-based SURF [11] divides the image into $N \times N$ patches and the keypoint detection, description and matching steps of object recognition are applied separately for each focused region. Thus, to define the visitation order of the regions, the saliency map of the image is computed and the patches are quicksorted in a descending order of average saliency value for each region.

When a given patch of the scene z is focused to be processed by patch-based SURF, keypoints are detected in z. To do this, a scale-space function is built and the determinants of a Hessian matrix are computed on the current patch and also on each of its 8 neighbours whose determinants have not been evaluated; this is important to guarantee that pixels sufficiently near the patch border are not incorrectly detected just because the determinants at neighbour patches were not computed yet. After this, keypoints are finally localized by suppressing non-maxima points in a $3 \times 3 \times 3$ neighbourhood (considering the current scale and adjacent scales above and below the keypoint's scale in the scale-space function) followed by interpolation, but considering only locations of the focused patch z. For more details about this process, one can see [11] and [25].

Although patch-based SURF and SURF have similar recognition accuracy, the advantage of patch-based SURF is that it is faster if the searched object is detected in the analysed scene. In this case, as it processes each image patch separately and iteratively, a faster recognition is expected in case the target is recognized before all patches are selected to be processed. It was verified that by using patch-based SURF it was possible to recognize objects using only 73% of the time require by SURF. In this experiment, the visitation order of each image patch was defined randomly, showing that just by processing the input scene iteratively it is possible to speed-up the search without decreasing the recognition accuracy. This and the other experiments cited in this paper were performed using the Object Recognition Database from PONCE research group [27]. More information about this experiment and about the proposed patch-based SURF can be found in [11].

B. Saliency detection: Background Laplacian Saliency

The saliency detector proposed herein, named Background Laplacian Saliency (BLS) [11], evaluates global and local saliency and combines them to achieve a final saliency map. Based on our saliency method proposed in [28], BLS computes saliency globally based on the estimative of the main background colour, that is computed as follows. Firstly, to disregard texture and noise, the image is convolved with a 5×5 Gaussian filter ($\sigma = 1.1$); by doing this, the image I_{gb} is generated. Then, Canny's edge detector [29] is applied and the Distance Transform is executed on the resulting edge image.

The distance transformed image (DT) is used to assign higher weights to pixels far from edges; by doing this, it is expected to bias the main background colour to the colours of large homogeneous regions. Thus, the Distance Transform Global Saliency (DTGS) is computed as:

$$DTGS(x,y) = ||I_\mu - I_{gb}(x,y)||, \quad (1)$$

where $I_\mu = [I_{\mu_L} I_{\mu_A} I_{\mu_B}]^t$ is the image mean feature vector in the Lab colour system, with each colour component of I_μ being defined based on the weighted average of the pixels in the DT image. Thus, each colour component in I_μ is evaluated as

$$I_{\mu_c} = \frac{\sum_{x=0}^{m-1}\sum_{y=0}^{n-1} I_{gb}(x,y,c)\cdot DT(x,y)}{\sum_{x=0}^{m-1}\sum_{y=0}^{n-1} DT(x,y)}. \quad (2)$$

where n and m are the height and the width of the image, $I_{gb}(x,y,c)$ is value of the pixel of I_{gb} at position (x,y) and at colour channel c, while DT stands for the distance transformed image.

Since $DTGS$ can disregard regions that, although being locally salient, do not stand out globally, the proposed method also uses the image's second derivative to detect salient regions locally. In order to do this, the Laplacian filter is applied at each colour channel of I_{gb}, and the arithmetic mean between L, a and b components is defined as the local saliency of each pixel. This is resumed as

$$LS(x,y) = g(x,y) * (1/3 \sum_{c=1}^{3} (L(x,y) * I_{gb}(x,y,c))), \quad (3)$$

with L representing the Laplacian kernel, while $*$ is the convolution operator and c is each colour channel used (L, a or b). The Gaussian kernel g is also used to detect saliency in a sparse way. Finally, we combine $DTGS$ and LS to achieve the final saliency map (BLS), as:

$$BLS(x,y) = (DTGS(x,y) + LS(x,y))/2. \quad (4)$$

Figure 3 shows the saliency maps produced by BLS and other four saliency detection algorithms using images from PONCE dataset.

Using BLS saliency maps to guide the search executed by patch-based SURF, the recognition time was decreased, in average, to only 53% of the processing time of classic SURF. Besides that, the visual search guided by BLS was faster than using other nine saliency detection algorithms. This experimental results and more information about BLS can be found in [11].

C. Integration Between Bottom-up and Top-Down Visual Attention

Differently from the bottom-up search, when a top-down mechanism is used it is aimed to direct the search towards regions that resembles the searched object. In our work, this is done by first, in a training phase, constructing a neighbourhood of the descriptors of the searched object. Then, in the test phase, when a scene is analysed, the descriptors that are presented in the neighbourhood are prioritized according to their Hamming distance to the descriptors of the searched object. In [23], we proposed an algorithm that generates the neighbourhood of a descriptor using the Depth First Traversal in a Tree representation of this same neighbourhood. Each descriptor of the neighbourhood is then represented in a probabilistic data structure called Bloom Filter [30].

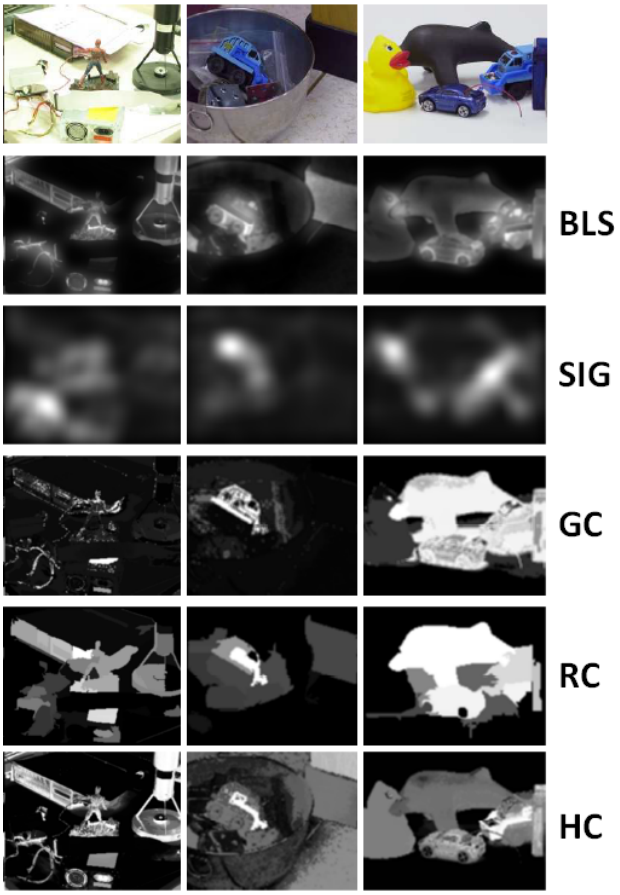


Fig. 3. Saliency maps produced by BLS [11] and other state-of-the-art saliency detection algorithms (SIG [15], GC [18], RC [13] and HC [13]).

When the bottom-up and top-down methods are integrated, the patch-based SURF is used to allow the processing of each image region separately and a saliency map is used to define the visitation order of the patches of the scene. However, once a given patch is selected to be processed, only the detection and description phases are executed. Then, instead of executing the matching with all descriptors of the selected patch, the top-down attention method is applied to prioritize the descriptors of the focused patch according to the similarity of each descriptor to the descriptors of the searched object. If the object is found using only prioritized descriptors, the search ends without processing neither the rest of the descriptors of the current patch, nor the remaining patches of the scene. Otherwise, if the object is not found, but if there is at least one matching between the descriptors of the scene and the searched object, a refinement of the matching is executed, when all remaining descriptors of the current patch

are considered in a new matching step.

In the case that no descriptor is sufficiently similar to the descriptors of the searched object, the algorithm considers that the target is probably not localized in the current region; then, the search proceeds to focus on the next more salient patch, and the descriptors are stored in memory to be used in a final matching step. This final matching is executed in the situation in which all patches are processed and the object is not recognized. Storing the descriptors not used in the matching step of a given patch to be matched after all patches are visited is important to guarantee that the top-down attention stage does not affect the recognition accuracy.

The advantage of integrating the bottom-up and top-down searches is that it defines the region to be focused based on the saliency of the scene (which is usually much faster to compute than recognition algorithms), without the need to execute any step of the recognition process. Moreover, the top-down attention, which is slower, but theoretically more accurate than bottom-up methods, makes it possible that the focused patch may be discarded before the execution of the matching step of the recognition if the descriptors of the focused patch are evaluated as not being sufficiently similar in relation to the descriptors of the searched object. Moreover, the top-down mechanism also decreases the processing time by prioritizing the descriptors that are more similar to the searched object.

In our Thesis, we have verified experimentally that, by using the integration between the bottom-up and top-down methods, the recognition time was decreased, in average, to only 44% of the processing time of classic SURF. In the worst case, our proposal requires 66% of the processing time of the Unified Visual Attention Model [24], which also integrates bottom-up and top-down attention mechanisms.

IV. Conclusion

Our Thesis explored the connection between visual attention and object recognition. The proposed bottom-up method uses a saliency map to guide the search for an object, so that locations that are more salient are prioritized. A major benefit of this approach is that it requires a very low fixed cost to define the visitation order of the scene, since the saliency map can be computed before all steps of the recognition process. Ten different saliency algorithms were tested in the proposed visual search and it was experimentally shown that seven of them outperformed a random search, showing the feasibility of using saliency maps to guide the visual search. Moreover, the proposed BLS outperformed the other saliency detectors in this task. On the other hand, it was also shown that the other three methods achieved equal or worse performance than a simple random search, what highlights the importance of using saliency algorithms of high accuracy and low processing time.

In the proposed top-down visual search, the visitation order of the scene is influenced by knowledge (which represents the set of neighbours descriptors of the searched object) and, consequently, by the goal of finding the target in the scene. Due to this, the search is directed towards regions of the scene that are similar to the characteristics of the searched object. Therefore, our top-down proposal is able to decrease the matching step of the recognition.

On the other hand, this approach has the drawback of requiring that keypoint detection and description steps of recognition are computed before the application of the top-down attention to prioritize scene descriptors in the matching step. This is different from the bottom-up search, in which the visitation order of the scene is defined before the recognition process, but, on the other hand, the search is directed based only on the saliency of the scene, without taking into account the characteristics of the searched object. Due to this, although the top-down method has proven to be efficient in prioritizing descriptors to the matching step, it achieves a worse processing time if compared to the bottom-up search.

Finally, when both methods are integrated, the recognition is faster than when both methods are used separately. this is explained by the fact that the definition of the visitation order of the regions can be executed by a saliency map, before any step of the recognition process and, in addition, if the top-down method does not consider that the characteristics of the focused patch are not sufficiently similar to the characteristics of the searched object, the attended patch is not further processed and other region is focused according to the saliency of the scene.

A. publications

As results of the research conducted for this Thesis, one book chapter, one journal and two conference papers were published. These publications are directly related to this research, and are listed below:

- Journal:
  - R. G. Mesquita and C. A. B. Mello, “Object recognition using saliency guided searching”, *Integrated Computer-Aided Engineering*, vol. 23, no. 4, pp. 385–400, 2016. (**Qualis A1, Impact Factor: 5.264**)
- Conferences:
  - R. G. Mesquita and C. A. B. Mello, “Segmentation of natural scenes based on visual attention and gestalt grouping laws”, in *IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, SMC 2013, United Kingdom, October 13-16, 2013, 2013, pp. 4237–4242. (**Qualis A2**)
  - R. G. Mesquita, C. A. B. Mello, and P. L. Castilho, “Visual search guided by an efficient top-down attention approach” in *IEEE International Conference on Image Processing*, 2016, pp. 679–683. (**Qualis A1, H5-index: 35**)
- Book Chapter:
  - R. G. Mesquita and C. A. B. Mello, “New developments in visual attention research”, in *Object Recognition Guided by Visual Attention Algorithms*,

1st ed. Nova Science Publishers, 2017, vol. 1, ch. 2, ISBN: 978-1-53612-374-6 [2].

In addition, the following works involving visual attention or visual perception were published during the course of this doctorate:

- Journals:
  - R. G. Mesquita, C. A. B. Mello and L. H. E. V. Almeida, " A New Thresholding Algorithm for Document Images Based on the Perception of Objects by Distance", *Integrated Computer-Aided Engineering*, vol. 21, no. 2, pp. 133-146, 2014. (**Qualis A1, Impact Factor: 5.264**)
  - R. G. Mesquita, R. M. A. Silva, C. A. B. Mello and P. B. C. Miranda, "Parameter tuning for document image binarization using a racing algorithm", *Expert Systems with Applications*, vol. 42, no. 5, pp. 2593–2603, 2015. (**Qualis A1, Impact Factor: 3.928**)
    - In this paper an improvement of our threshold algorithm based on visual perception was presented. It achieved the first place in the H-DIBCO (Handwritten Document Image Binarization Contest) [31].
- Conference:
  - R. G. Mesquita and C. A. B. Mello, "Finding Text in Natural Scenes by Visual Attention", in *IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, SMC 2013, United Kingdom, October 13-16, 2013, 2013, pp. 4243–4247. (**Qualis A2**)

ACKNOWLEDGMENT

This research was partially sponsored by CNPq under Grant 141190/2013-2, by INES and by CAPES.

REFERENCES

[1] J. K. Tsotsos, *A Computational Perspective on Visual Attention.* MIT Press, 2011.

[2] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[3] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.*, vol. 7, no. 1, pp. 6:1–6:39, Jan. 2010.

[4] R. Mesquita and C. Mello, "Reconhecimento de instâncias guiado por algoritmos de atenção visual," Ph.D. dissertation, Centro de Informática, Universidade Federal de Pernambuco, February 2017.

[5] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[6] D. Kim, S. Rho, S. Jun, and E. Hwang, "Classification and indexing scheme of large- scale image repository for spatio-temporal landmark recognition," *Integrated Computer-Aided Engineering*, no. September 2015, 2015.

[7] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Coding Visual Features Extracted From Video Sequences," *IEEE Transaction on Image Processing*, vol. 23, no. 5, pp. 2262–2276, 2014.

[8] D. Feng, J. Yang, and C. Liu, "An efficient indexing method for content-based image retrieval," *Neurocomputing*, vol. 106, no. 0, pp. 103 – 114, 2013.

[9] J. Hou, Z. Chen, X. Qin, and D. Zhang, "Automatic image search based on improved feature descriptors and decision tree," *Integrated Computer-Aided Engineering*, vol. 18, pp. 167–180, 2011.

[10] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. "Cambridge University Press, ISBN: 0521540518", 2004.

[11] R. G. Mesquita and C. A. B. Mello, "Object recognition using saliency guided searching," *Integrated Computer-Aided Engineering*, vol. 23, no. 4, pp. 385–400, 2016.

[12] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 733–740.

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

[14] B. Jiang, L. Zhang, H. Lu, C. Yang, and M. H. Yang, "Saliency detection via absorbing Markov Chain," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1665–1672.

[15] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[17] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection." in *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1597–1604.

[18] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, 2013, pp. 1529–1536.

[19] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient Object Detection: A Survey," *arXiv*, vol. 1411.5878, 2014.

[20] ——, "Salient Object Detection: A Benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[21] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 438 – 445.

[22] Y. Liu, Q. Jian, X. Zhu, J. Cao, and H. Li, "Saliency Detection Using Two-stage Scoring," in *Int. Conference on Image Processing*, 2015, pp. 1–5.

[23] R. G. Mesquita, C. A. B. Mello, and P. L. Castilho, "Visual search guided by an efficient top-down attention approach," in *International Conference on Image Processing*. IEEE, 2016, pp. 679–683.

[24] S. Lee, K. Kim, J.-Y. Kim, M. Kim, and H.-J. Yoo, "Familiarity based unified visual attention model for fast and robust object recognition," *Pattern Recogn.*, vol. 43, no. 3, pp. 1116–1128, Mar. 2010.

[25] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[27] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *International Journal of Computer Vision*, vol. 66, no. 3, pp. 231–259, 2006.

[28] R. G. Mesquita and C. A. B. Mello, "Segmentation of natural scenes based on visual attention and gestalt grouping laws," in *IEEE International Conference on Systems, Man, and Cybernetics, Manchester, SMC 2013, United Kingdom, October 13-16, 2013*, 2013, pp. 4237–4242.

[29] J. Canny, "A computational approach to edge detection." *IEEE transactions on pattern analysis and machine intelligence*, vol. 8, no. 6, pp. 679–698, Jun. 1986.

[30] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[31] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "ICFHR 2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014)," in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Sep. 2014, pp. 809–813.

[2] https://www.novapublishers.com/catalog/product_info.php?products_id=62836