A Novel Approach to the Extraction of Multiple Salient Objects in an Image

Srikanth Muralidharan*, Arun Balajee Vasudevan*, Chintapalli Shiva Pratheek*, Shanmuganathan Raman†
*Electrical Engineering, Indian Institute of Technology Jodhpur
Email: srikanth@iitj.ac.in, arunbalajeev@iitj.ac.in, shiva@iitj.ac.in
†Department of Electrical Engineering, Indian Institute of Technology Gandhinagar
Email: shanmuga@iitgn.ac.in

***Abstract*—This paper describes a novel approach for extraction of multiple objects from a given image of a natural scene. In the proposed approach, multiple objects are extracted by the application of saliency detection on the image. We use two distinct approaches for object extraction. One approach uses superpixels on the saliency map. Then the intensity of saliency map in each superpixel is used to compute distance between the centres of superpixels. These act as constraints to extract the objects from the image. The other approach is the application of Active Contour model on the saliency map and estimating a bounding box on the intermediate binary image result extracts the objects in the image. The approach is unique in its way of extracting objects from a scene containing multiple objects as it does not use extensive image search like the existing algorithms and therefore leads to a fast and simple extraction.**

I. Introduction

Most of the available digital images contain both redundant and useful information, which occurs in form of static and dynamic objects. In order to extract, many computer vision researchers are trying to develop algorithms which can perform as good as the human visual system (HVS) perception of objects. These algorithms, if implemented efficiently to extract the significant objects, would find applications in various practical problems of computer vision. Examples include detection of objects from surveillance videos, enhancement of objects in images captured at different times of a day, and video synopsis.

We propose to extract useful information namely the salient objects from a given image. Our approach differs from the other existing algorithms as we do not perform extensive search in the image for objects. The proposed approach involves the use of visual saliency, active contours, and superpixels as building blocks to achieve the desired objective. The reason for choosing visual saliency for object extraction is that it closely imitates the HVS perception and detects the information relevant to the user.

We would like the proposed approach to compete with other approaches which first perform the object discovery followed by the localization [5], [6]. Objects in images are discovered and recognized by comparing models after learning in an unsupervised setting[7], [8]. Fig. 1 shows the comparison of the proposed approach against the other state-of-the-art methods. As shown, the proposed approach matches the performance of the other state-of-the-art methods in localizing the object (vehicle). Our approach performs the extraction of multiple objects in an image directly without the need of any image datasets for learning.

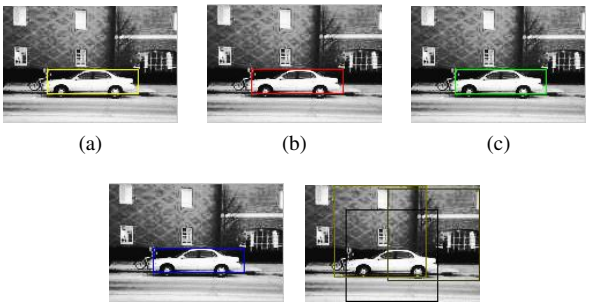


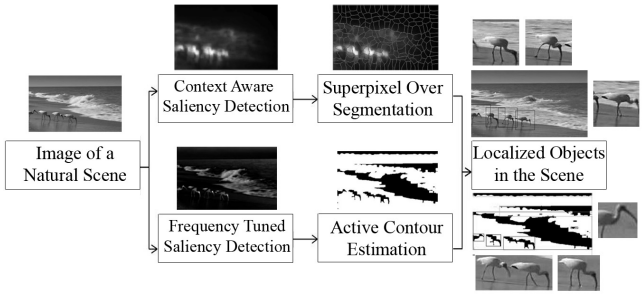
Fig. 1: Comparison of (a) Efficient Subwindow Search (ESS) [1], (b) Improved ESS [2], (c) ESS of Bentleys algorithm [3], (d) Alternating ESS [4], and (e) Multiple object extraction using the proposed approach.

The primary contributions of this paper are 1) Extraction of objects from an image of a natural scene. 2)The proposed algorithm is shown to be robust to the variations in depth of an object and does not involve extensive search for objects. 3) We are not interested in segmenting out the exact boundary of an object while we want to spatially extract it approximately from the image. The present approach will improve existing segmentation techniques.

In Section 2, a brief description about the related research done previously is provided. Section 3 contains a description about the design of proposed algorithm for multiple object extraction. In section 4, results of applying the proposed algorithm on several dynamic scene data sets are described. Section 5 provides the conclusion of the work. Section 6 discusses some of the challenges and directions for future research.

II. Related Work

Since the origin of scene classification, many approaches have been developed for selective extraction of objects. One of the latest automated approach for object extraction was given by Yu *et al.* in [9] about object extraction using complimentary saliency maps. Similarly, an attempt was made to classify events in static images by integrating scene and object categorization [10]. Object Localization can be efficiently done by Efficient Sub-window Search [1] which performs localized detection and image retrieval for classifiers. Additionally,



(a)

Fig. 2: The Proposed Approach

object localization can be performed by Bayesian correlation which is a synthesis of cross correlation matching [11] that takes care of occlusion and versatile enough to work with 3D pose changes. There are different efficient object search algorithms developed [2], [1], [13]. These algorithms use a subwindow to search and identify the object location. These algorithms are computationally efficient than previous object search algorithms and also identify the object location in the image.

Frequency tuned Salient region detection by Achanta *et al.* [14] is one of the standard saliency detection approaches which is based on frequency domain analysis and exploiting colour and luminance characteristics. The other prominent saliency detection is context aware saliency of the foreground objects. The rules that this algorithm use are inspired from psychological experiments which were conducted to find the features that humans use to find a salient object [15]. Many algorithms apart from visual saliency have been developed which can localize interesting objects from a scene such as energy function optimization approach [16], [17] and edge-linking methods [18] which connects subset of fragments generated by the edge-detection techniques.

Superpixels can be obtained by over-segmenting an image [19] which clearly demarcates the object edges from the background or other occluding objects. Superpixel approach can be seen in many state-of-the-art algorithms as one of the primary components for its efficiency [20]. Aggregating neighbourhoods of superpixels, Fulkerson *et al.* proposed a robust region classifier by object localization and segmentation [21]. Active Contours is an efficient way of outlining an object from an image. The major work in the Active Contours is active contours without edges [22]. We use this technique as a tool in the proposed approach.

III. Proposed Approach

The existing object localization algorithms involve a sub-window object search for object detection. The overview of proposed approach is shown in Fig. 2. The proposed approach involves two independent approaches that localizes multiple objects present in the scene. In both the methods, first we extract the saliency map of the input image and do the segmentation by superpixels to localize salient objects. The second method involves active contour to localize the salient objects. The main difference in the proposed approaches over the previous localization algorithms is that the complexity of the proposed approach is independent of the size of an image. The other algorithms are quadratic or bi-quadratic over the number of pixels in terms of complexity. Moreover, the proposed approach extracts the potentially distinct salient objects, in addition to localizing them in an unsupervised framework.

HVS can differentiate various objects in a scene by focusing attention on the attributes of objects such as its contrast, colour, and motion. Consider a video of a natural scene. Unsupervised learning with the help of the appropriate feature vector explicitly demarcates different regions of a scene into dynamic and static regions. In most cases, we observe multiple salient sub-regions in a scene, especially in the static regions. These stationary objects devoid of motion can be localized by the saliency. Visual saliency has been shown to be coherent with HVS perception. It is shown in this work that object localization can be accomplished by saliency based approach.

We formulate an approach to extract multiple object regions separately from the given image of a scene. We use context aware saliency estimation which aims at detecting the image regions that represent the scene rather than identifying fixation points or dominant objects [15]. We use two separate methods for sub-image localization from saliency image. Amongst all the saliency detection algorithms, context based saliency algorithm provides the best spatial cover for the foreground object and therefore is well suited to superpixel based method, which relies on spatial configurations of the input [15]. On the other hand, Active Contour model uses frequency tuned saliency detection that marks well defined boundaries for the detection of salient objects in the scene [14], in contrast with context based saliency method. This is exploited by Active Contour method to extract the shape of the object.

A. Superpixel Method

Consider the saliency map extracted from a given image. The first step involves over-segmentation using superpixels[1] from the saliency map [15]. We initialize number of distinct segments, typically to a value between 10 and 20. After over-segmenting the saliency map into distinct segments, we take the average intensity value of each superpixel. We further refine the number of interesting superpixels or sub-regions by forcing the cumulative superpixel intensity values to zero if the average intensity values are below a certain threshold, ϵ.

If the average intensity of saliency map is higher as in the case of a dense scene, we keep ϵ to be equal to the average intensity value of saliency map. Having a higher multiple ($1.5\times$) of average saliency map intensity as ϵ would result in elimination of detection of certain salient object regions. If the average intensity value is less (a sparse scene), ϵ should be kept at much higher multiple ($2\times$) of the average saliency map intensity value. Lower value of ϵ leads to the generations of many redundant sub-images. We use a distance constraint to decrease the number of redundant images, that is, half perimeter distance between centres of each superpixel as the distance measure. Let C_x, C_y denote x and y co-ordinates of the centre of superpixels and d_{ij} denote the distance between

[1] http://www.cs.sfu.ca/ mori/research/superpixels/

the centres of superpixels i and j. The distance d_{ij} is given by the sum of absolute difference (SAD) equations (1-3).

$$d_{ij} = |C_x(i) - C_x(j)| + |C_y(i) - C_y(j)| \quad (1)$$

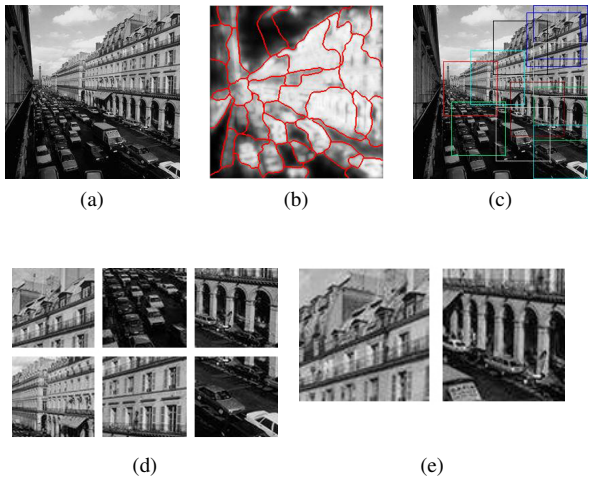
$$C_x(k) = \frac{1}{N(k)} \sum_{m \in k} x(m) \quad (2)$$

$$C_y(k) = \frac{1}{N(k)} \sum_{m \in k} y(m) \quad (3)$$

Here $N(k)$ denotes number of pixels inside superpixel k. Value of d_{ij} higher than a threshold, δ makes the two sub-regions to be considered as distinct salient objects. The value of δ is decided by the average of intensity value of the saliency map. For a high value of average intensity value of the saliency map, δ will be set a lesser value (typically less than 10 pixels). This results in a single sub-image from this region and avoids redundancy. A sub-image is finally extracted centred at each superpixel which follows the defined constraints.

B. Active Contour Method

The alternate approach involves extracting multiple objects using active contours on the saliency map. Saliency region detection [14] leads to determination of saliency map with well-defined boundaries of salient objects. These boundaries can be used to segment the objects by application of active contours [22] on the saliency map as shown in the Fig. 8(d).



(a) (b) (c)

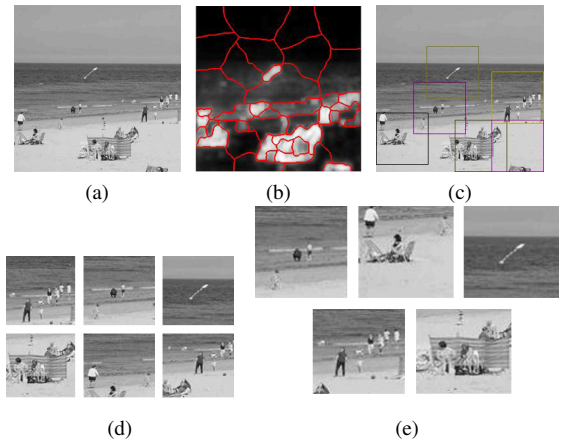
(d) (e)

Fig. 3: (a) An urban street scene, (b) Superpixel image of saliency map, (c) Extracted segments represented by bounding box, Result of Keeping ϵ at 0.75$\times$ (d), and at 1.5$\times$ (e), the average saliency map intensity value.

We choose the initial contour for active contour to cover the entire scene to capture objects at any place. We choose small circles placed in the entire scene to be the contour as salient objects can be small in a natural scene. Although the image boundaries are smooth and noisy, the location of boundaries are well detected by active contours and gives a binary image of the scene highlighting the objects. We use the connected component labelling where subsets of binary connected components using 8- neighbourhood are uniquely labelled which helps in object extraction. We make use of the bounding box to extract out object parts from the scene frame. Active Contour has advantage in the detection of large number of objects in the scene while we make a bound for the superpixel based approach with the initialisation of the number of divisions. Active Contour model also gives the apposite shape of an object delineating the object outline. Active Contours help to even detect the vacant portion within the object in the 2D image.

IV. Results

In order to see the performance of proposed approach, we took a collection of images of varied complex scenes such as streets, kitchen, forest, coast, industry, etc. [23].



(a) (b) (c)

(d) (e)

Fig. 4: (a) A sparse coast scene, (b) superpixels image of saliency map, (c) Extraction represented by bounding box, Result of Keeping ϵ 0.5 $\times$ (d), and 2.5 $\times$ (e), of the average saliency map intensity value.

To illustrate the effect of ϵ based redundancy removal, consider a dense scene as shown in the Fig. 3(a). The saliency map corresponding to the scene was obtained using Context aware Saliency algorithm [15]. The average intensity of the saliency map is 146. The segmented superpixel map of saliency map with 20 superpixels is shown in Fig. 3(c). ϵ is assigned two values: First at 0.75 and the second at 1.50 times the average intensity value, keeping other parameters constant. The 80×80 bounding box is then used to extract superpixels with non-zero intensity values. Fig. 3(d) and Fig. 3(e) show the set of salient regions extracted using the first and second threshold values. Here we note that the scene contains different salient objects like buildings with distinct salient regions, cars, vans, and a distant pillar. Setting a lower ϵ extracts these salient regions without any redundancy as shown in Fig. 3(d). As shown in Fig. 3(e), the number of salient objects extracted will be less than the actual number of salient objects present in the scene when ϵ is set high. Consequently, we lose potentially important salient regions by keeping ϵ high in a scene having dense distribution of objects.

Consider another scene which is sparse as shown in Fig. 4(a). The saliency map corresponding to the scene was obtained using Context aware Saliency algorithm [15]. The average intensity of the saliency map is 58. The segmented

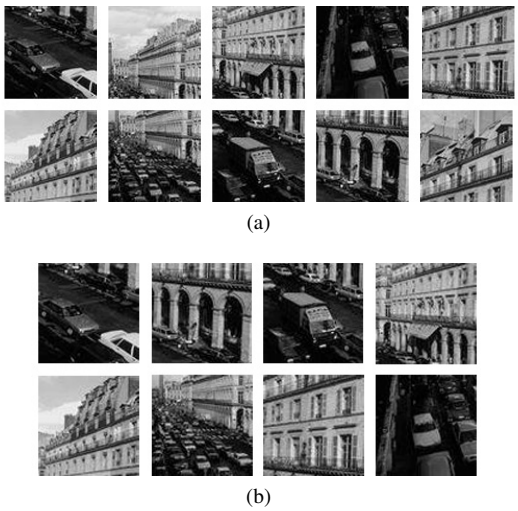


Fig. 5: (a) Result of thresholding the saliency map with δ as 10 pixels, (b) 60 pixels.

superpixel map of saliency map with 20 superpixels is shown in Fig. 4(c). The superpixel map is made to have two values of ϵ: First at $0.5(= 29)$ and the second at $2.5(= 145)$ times the average intensity value, keeping other parameters constant. The 80×80 bounding box is then used to extract superpixels with non-zero intensity values.

Fig. 4(d) and Fig. 4(e) show the set of salient regions extracted using different threshold values. Here we note that the scene contains sparsely distributed salient objects. Setting a lower ϵ extracts these salient regions separately as shown in Fig. 4(d). In Fig. 4(e), the salient sub-images are obtained at higher value of ϵ. As shown in the figures, we observe that salient objects extracted with higher and lower ϵ are almost the same. We can also observe that there is redundancy while extracting the salient objects. There is no difference in redundancy removal with a higher value of ϵ.

To illustrate the effect of distance constraint, consider again the urban street scene shown in Fig. 5. When δ is equal to 10, the sub-images extracted by the approach is obtained in Fig. 5(b). When δ is equal to 60 with other parameters constant, the sub-images extracted is shown in Fig. 5(b). From the figures shown, we observe that the number of salient objects extracted decreases as δ increases. Here, we fail to extract some of the salient regions as we further increase δ.

We have a coast scene with birds on the beach as shown in the Fig. 6(a). We used frequency tuned saliency region detection which provides well defined boundaries for the salient objects as in Fig. 6(b). Application of the Chan-Vese Active Contour model without edges on the saliency map [22] separated out the object as represented in the binary image Fig. 6(d). We used the initial contour as shown in Fig. 6(c) with small circles that uniformly fill the entire space. The initialisation of the number of iterations is made to 400 which is a significant part of Active Contour algorithm. We initialized the length term to be 0.005 which adjusts the curvature of the contour for each iteration for the segmentation. Birds and beach are extracted out from the scene using the bounding

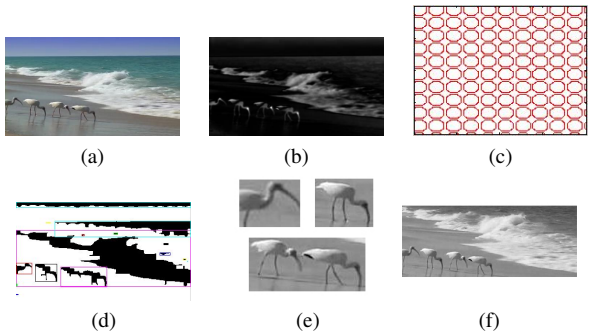


Fig. 6: (a) A coast scene, (b) Frequency tuned saliency map of the scene, (c) Inital contour for Chan-Vese algorithm, (d) Binary Image model of Active Contour model on saliency map, (e) and (f) are objects extracted applying bounding box approach on binary image avoiding non-salient objects by removing objects with low average intensity ratio.

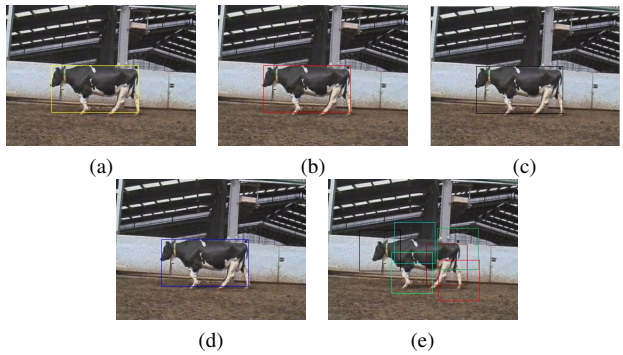


Fig. 7: Effect of saliency map on extraction: (a) ESS algorithm [1], (b) Improved ESS [2], (c) ESS of Bentleys algorithm [3], (d) Alternating ESS [4], (e) output of the proposed approach.

box approach Fig. 6(d). Though other objects are also detected by the bounding box approach, they are not considered to be extracted as they have a low average value of the saliency intensity over the region.

Fig. 7 shows images of cow and car where the outputs of previous algorithms are shown in Fig. 1(a, b, c, d,), Fig.7(a, b, c, d) and the output from the proposed approach are shown in Fig. 1(e) and Fig. 7(e). We observe that the output of the proposed approach shows object extraction at various locations of cow's body because of the saliency map that assigned bright intensity to the portions of cow. In the car image Fig.1(e), we observe multiple objects being detected in the scene.

Fig. 8(a), and 8(c) shows the plot of fraction of redundancy and undetected salient objects against several values of ϵ, keeping δ constant at 50 pixels for the coast scene and 40 pixels for the city scene. Fig. 8(b), and 8(d) shows the plot of fraction of redundancy and undetected salient objects against several values of δ, keeping ϵ to their respective average intensity values. We observe from the plot that for the city scene, the proportion of undetected salient regions increases at low values of ϵ and δ. On the other hand, for the coast scene, it is observed that the proportion of redundant salient regions detected is high at low δ, independent of the value of ϵ (Fig.

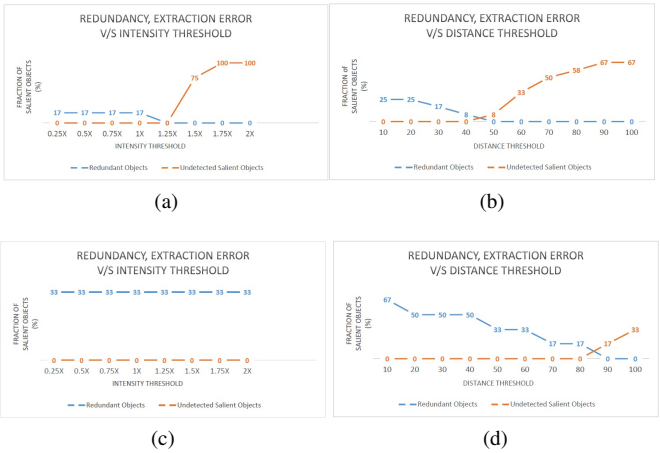


Fig. 8: Redundancy, detection error plots for fixed δ and several values of ϵ for (a) city scene, and (c) coast scene, Redundancy, detection error plots for fixed ϵ and several values of δ for (b) city scene, and (d) coast scene.

8(c)). For this scene, we observe from this plot that redundancy decreases gradually with the increase of δ at a fixed ϵ. These plots generalize the observations made for different scenes at specific values of ϵ, and δ in the early part of this section to a range of values.

V. Conclusion

This paper proposes two different methods of multiple object extraction from an image. First, objects are obtained using context aware saliency detection and superpixel over-segmentation. In this method, both ϵ, and δ depend on the scene. For a dense scene with close objects, both the thresholds should be lower. For a sparse scene with scattered objects, they should be set high. Setting the appropriate values result in efficient extraction of salient sub images.

The Active Contour techniques on the saliency map also gives multiple objects with no bounds on the number of objects in the scene. Active Contour produces better results as compared to superpixel based method when there is a large, single object present. On the other hand, superpixel based method produces better results when the distance between salient objects is very small and also in the cases where the object is occluded. Thus, two approaches complement each other in such cases and together extract the entire set of salient sub-regions from the image.

VI. Future Work

An automated framework for estimation of ϵ, and δ for different scenes has to be developed. The proposed approach largely relies on the output of saliency map. A new framework needs to be developed for the cases where saliency map fails to detect salient regions, like in the cases where objects which have almost same color as background. The proposed approach shall be applied in classification algorithms for labelling the scenes with multiple objects present. Extracted objects shall be used for the object classification in a scene. It has scope for the image compression in video avoiding redundancy of storing the same object multiple times. In addition, we also aim to do a comprehensive evaluation of our method on more challenging datasets.

References

[1] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *IEEE CVPR*, 2008, pp. 1–8.

[2] S. An, P. Peursum, W. Liu, and S. Venkatesh, "Efficient algorithms for subwindow search in object detection and localization," in *IEEE CVPR*, 2009, pp. 264–271.

[3] J. Bentley, "Programming pearls: algorithm design techniques," *Communications of the ACM*, vol. 27, no. 11, pp. 1087–1092, 1984.

[4] S. An, P. Peursum, W. Liu, S. Venkatesh, and X. Chen, "Exploiting monge structures in optimum subwindow search," in *IEEE CVPR*, 2010, pp. 926–933.

[5] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *IEEE ICCV*, vol. 1, 2005, pp. 370–377.

[6] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *IEEE CVPR*, vol. 2, 2006, pp. 1605–1614.

[7] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 284–302, 2010.

[8] Y. J. Lee and K. Grauman, "Object-graphs for context-aware category discovery," in *IEEE CVPR*, 2010, pp. 1–8.

[9] H. Yu, J. Li, Y. Tian, and T. Huang, "Automatic interesting object extraction from images using complementary saliency maps," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 891–894.

[10] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *IEEE CVPR*, 2007, pp. 1–8.

[11] J. Sullivan, A. Blake, M. Isard, and J. MacCormick, "Object localization by bayesian correlation," in *IEEE ICCV*, vol. 2, 1999, pp. 1068–1075.

[12] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *IEEE ICCV*, 2009, pp. 817–824.

[13] J. L. Bentley, *Programming Pearls, 2/E*. Pearson Education India, 2000.

[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE CVPR*, 2009, pp. 1597–1604.

[15] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 1915–1926, 2012.

[16] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level set evolution without re-initialization: a new variational formulation," in *IEEE CVPR*, vol. 1, 2005, pp. 430–436.

[17] N. Xu, N. Ahuja, and R. Bansal, "Object segmentation using graph cuts based active contours," *Computer Vision and Image Understanding*, vol. 107, no. 3, pp. 210–224, 2007.

[18] S. Wang, T. Kubota, J. M. Siskind, and J. Wang, "Salient closed boundary extraction with ratio contour," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 4, pp. 546–561, 2005.

[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[21] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *IEEE ICCV*, 2009, pp. 670–677.

[22] T. F. Chan, B. Y. Sandberg, and L. A. Vese, "Active contours without edges for vector-valued images," *Journal of Visual Communication and Image Representation*, vol. 11, no. 2, pp. 130–141, 2000.

[23] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *IEEE CVPR*, vol. 2, 2005, pp. 524–531.