Benchmarking Saliency Detection Methods on Multimodal Image Data

Hanan Anzid[1,2](✉), Gaetan Le Goic[2], Aissam Bekkari[1], Alamin Mansouri[2], and Driss Mammass[1]

[1] IRF-SIC Laboratory, Faculty of Science, Agadir, Morocco
hanananzid@gmail.com

[2] LE2I Laboratory, University of Burgundy-Franche-Comte, Dijon, France

Abstract. Saliency detecmage processing. Most of the work is adapted to the specific application and available dataset. The present work is about a comparative analysis of saliency detection for multimodal images dataset. There were many researches on the detection of saliency on several types of images, such as multispectral, natural, 3D and so on. This work presents a first focused study on saliency detection on multimodal images. Our database was extracted from acquisitions on cultural heritage wall paintings that contain four modalities UV, IR, Visible and fluorescence. In this paper, the analysis has been performed for many methods on saliency detection. We evaluate the performance of each method using NSS similarity metric. The results show that the best methods are [16] for visible modality, and [20] for UV, IR and fluorescence modalities.

Keywords: Saliency map · Multimodal images

1 Introduction

Multimodal data is a challenging area in image processing, that gained much attention of many researchers due to the benefit it provides to solve a given problem with some various available information. In this area the problem of visual saliency detection is a profound challenge task.

Several researches provide a different definition of saliency detection. Classically it aims to predict what is likely to catch the human attention [7,15,28], which can benefit a large variety of real-world applications, e.g. image segmentation [13], object recognition [17,24,26], visual tracking [25], image/video compression [30], object detection [24].

The saliency detection methods can be divided into Bottom-Up saliency detection and Top-Down saliency detection problem. The Bottom-Up saliency detection methods require no prior knowledge of the image and they are based on a visual features attention, such as intensity, colour, orientation of edges and direction of movement, which [14,15] is an example implementation of this model. The Top-down saliency detection methods are task oriented and demand

A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 11–18, 2018.
https://doi.org/10.1007/978-3-319-94211-7_2

prior knowledge about the context to identify the salient regions, relying on high-level visual tasks, such as searching a specific object that consciously catches an observers goal and attention e.g., faces, human, cars, etc. One implementation of this category is the work by Torralba et al. [2,21–23].

In this paper, we are using four modalities of images, UV, IR, UVF and visible about the same scene to emphasize specific information for a later classification. The underlying idea is that the saliency region is distinctive according to the modality used. So, we are analysing and comparing the various saliency detection methods on this images dataset, in order to choose the method that gives the most accurate results for further use. We are evaluating the performance of each method using NSS similarity metric. The rest of the paper is organized as follows. In Sect. 2 we propose the literature survey of used methods. Data and Experimental results are presented in Sect. 3. Section 4 discusses results and concludes this work.

2 Literature Survey

Globally, the Bottom-up saliency detection methods are clustered into four main categories [28]:

Space-Based Model: it is based on biologically motivation features selection, followed by computing centre-surround differences of selected primitives (colour, edge, intensity). The model of Itti [14,15] is the mile-stone example.

Frequency-Based Model: this model uses spectral domain instead of a proper image. The well-known model of this category is SR [29] that is based on a notion of spectral residual which uses log spectrum of the image derived from the averaged Fourier spectrum.

Object-Based Model: this model relies on the theory that attention is allocated through the perceptual object created by Gestalt rules, such as, proximity, similarity, closure, continuation and so on. The model of Bruce et al. presents a pioneering work of this kind which used the principle of maximizing information [5].

Graph-Based Model: this model is based on a notion that information is more powerful behind attentive sampling. GBVS presents the classical model [8].

Whereas a top-down method is generally based on the bottom-up process, and consists of two main steps, dictionary learning and saliency computation [9]. But it still receives less research attention due to the complexity to mimicking a high level of visual attention cognitive process [6].

In this paper, we present a review of the best-known methods for saliency detection in both Bottom-up and Top-down methods used in this work to detect saliency map on multimodal images.

In [15] Itti proposes an earlier and classic model reflecting what likely catches a human eye attention. This model is widely used to calculate saliency region in a given image, and is considered as guideline for several works as a tool for comparison purpose. The authors used the features of colour, intensity and orientation as the primitive information. Extracted features are processed in parallel way across several spatial scales after a linear filtering. Then conspicuity maps are computed through center surround. The mean value of this maps are computed to get a final saliency map.

The [4] presents a scalable framework that relies on object-based method for scene classification using decision trees. The authors proposed an automatic ROI detection to provide a saliency guided object discovery, which is used as foreground markers for object segmentation by graph cut. Object classification step is then performed using the bag-of-features with SIFT descriptors and support vector machine SVM. In this step authors integrate a special pyramid matching to improve object accuracy. A scene classification is then carried out using a decision tree constructed by labelled training images.

In [11], Hou et al. suggest an algorithm to separate foreground from background part of the image. The segregation is performed by using the sign function of Discrete Cosine transform (DCT) of an image. Then the saliency map is generated by computing the inverse of Discrete Cosine transform (IDCT) to extract the foreground part of image which is usually considered the most prominent. The approach is tested for both RGB and CIELAB colour channels and CIELAB produces better results.

Zhang et al., present in [30] a new Boolean Map saliency model. The proposed model relies on surroundedness cue as a Gestalt principle which is enclosured to topological relationship between the figure and the ground of the image, in order to apply invariant separation to the transformation. In this model an image is characterized by a set of Boolean maps obtained from selected features channels. On these maps, a binary image technique is applied to highlight regions contour, in order to compute an attention map. Then a linear combination of attention map is applied to obtain a mean attention map that can be used with some post-processing to provide the saliency map.

In [7], the authors propose a new type of salience based on detection of either fixation points or dominant object based on local and global surrounding information. This new algorithm relies on four principals observed in the psychological literature: **local low-level considerations** such as contrast and colour, **global considerations** which aim to keep features that deviate from the norm and gives with low-level an immediate context, **visual organizational rules** that organized the form by regrouping the salient pixels together instead of being spread all over the image, and **high-level factors** such as recognized objects or face detection that are applied as post-processing.

Rahtu et al. [20] propose a new segmentation method based on saliency object for image and video sequence. The authors used local feature contrast in illuminance, colour and motivation information controlled by smoothing parameter with a statistical framework in order to provide saliency map measure. The saliency map measure is combined with a conditional random field (CRF) in order to achieve segmentation purpose.

Le Moan et al. propose in [16] a new method for saliency detection inspired by Itti model for a purposes of image comparison, visualization and interpretation. This model is achieved in three steps. The first step is **Measure, features and center-surround comparisons** in which centers and surrounds are defined and then compared using four comparison measures as Euclidean distance (L2-norm), CIELAB trichromatic values, the angle between pixels and the Gabor-filtered image of the first principal component. Reflectance, principal components and Gabor-filtered pixels are the features used by this measures are then compared in turn using Euclidean distance or angle. The second step is **Normalization** that aims to concentrate lightness into less key location on a fused map. And the last step is **low-dimensional images**.

In [29] Hou et al. propose a novel model that used the spectral domain of the image instead of the proper image. The main of the algorithm is to use spectral residual generated from the difference logarithmised transformation and generalised shape of the image, in order to remove redundant information and emphasize only the new information. The saliency map is generated by applying an Inverse Fourier Transform on spectral residual and the values are squared and Gaussian smoothed.

3 Data, Experiments and Results

3.1 Data

This work evaluates best-known methods for saliency detection on multimodal images for the first time. The database used is extracted from acquisitions on cultural heritage wall paintings for detecting the original painted area and the repainted one. This set contain four modalities UV, IR, Visible and fluorescence of the same area or object. The use of four modalities together provides a benefit of various information instead of a single information source presented by one single image that gives only a partial distorted information [3]. An example of this set of images for the same area is presented on Fig. 1.

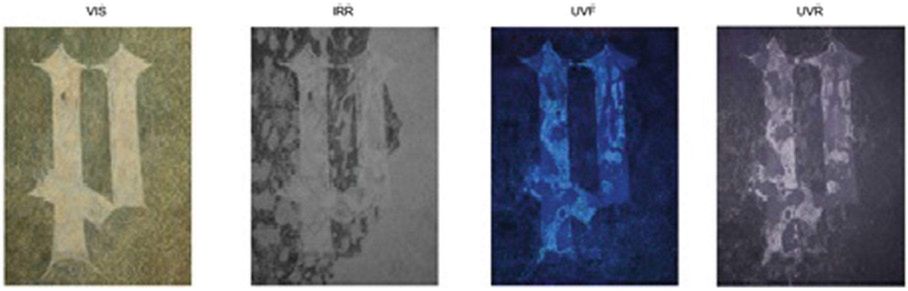

Fig. 1. Multi-modal images of the same are

3.2 Experiments and Results

The presented area in Fig. 1 as an example of the test contains trace of original paint and repainted region. Each modality emphasizes a part of these two classes. In this section, the results are presented which are obtained by applying saliency detection methods presented in the previous section. The evaluation is performed using NSS measure.

The graph-based visual saliency [8] is used in this work as a generator of fixation map, referring to the work [27] in which the author choose GBVS to obtain the fixation map for their dataset and proposed a new way to generate it by combining the result of multiple models. This model is considered more reliably than the standard algorithms.

We provide an exhaustive comparison of all methods cited in the survey section. Figure 2 shows a visual comparison of the different methods.

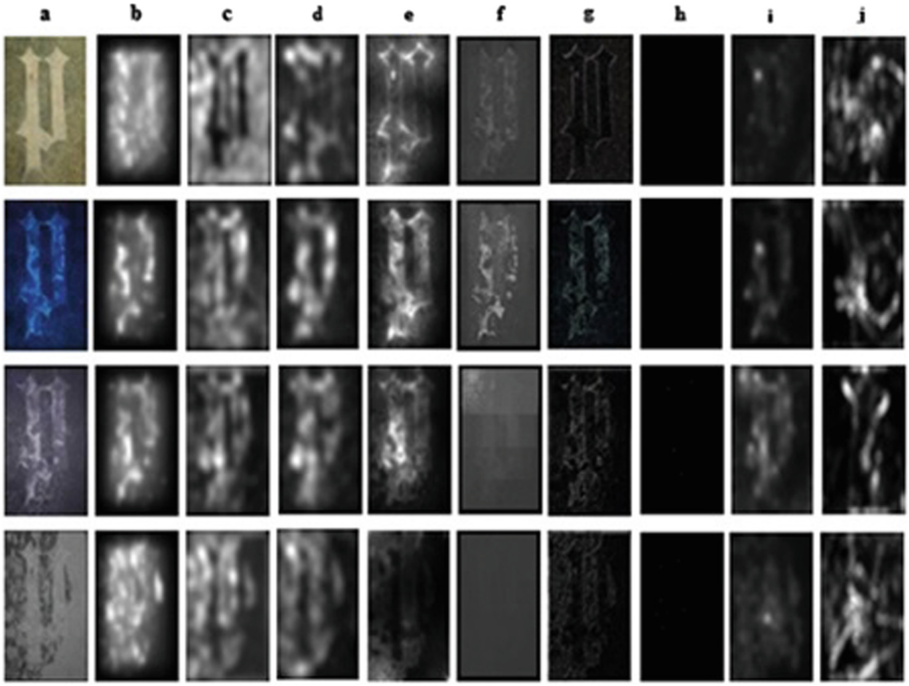

Fig. 2. Saliency map comparisons in Multimodal image database (a) original image. (b) Fixation map. (c) [14]. (d) Signature Saliency method [11]. (e) Context aware [7], (f) Segmenting salient objects from images and videos [20], (g) Spectral saliency [16], (h) ROI [4], (i) BMS [30], (j) Spectral residual method [29]

We evaluate the performance of each method measuring its Normalized Scanpath Saliency (NSS) metric [19] that quantify the saliency map values at the eye fixation map location. NSS is defined to have zero mean and unit standard deviation. The results summarized in Table 1 show that Itti model [15] performs better in the UVF and IRR modalities. Context-based aware [7] shows better score in UVF modality. Segmenting salient objects from images and videos [20] gives a high score in VIS, UVF and UVR modalities and Spectral saliency detection [16] performs better in visible modality.

4 Discussion and Conclusion

We performed a comparative analysis of saliency detection methods on a real multimodal images VIS, UVR, UVF and IRR, and compare the power of the resulting maps to the fixation map that is generated by existing eye-fixation prediction models [27]. The methods which are used here are both Bottom-up and Top-down methods.

Table 1. NSS metric of compared methods

Methods	[14]	[4]	[11]	[30]	[7]	[20]	[16]	[29]
VIS	0.7200	0.0523	0.6000	−0.5739	0.4200	**1.0100**	**5.1600**	0.2800
UVF	**1.2000**	0.09451	0.9600	0.6900	**1.0000**	**1.1720**	−17.960	−0.6500
UVR	0.5100	0.06174	0.5700	0.0700	0.4700	**1.0100**	−37.923	−0.3300
IRR	**1.0600**	0.02176	0.2400	0.1900	0.4200	0.4400	−0.2800	0.0572

Each modality contains information about original and repainted region of the same area, and each one emphasizes a specific region either in term of color, contrast, intensity, reflectance, texture or luminance, which can clearly catch a human attention. The NSS metric is widely used for saliency comparison, and shows through a computed score that many used methods are not perfectly suited for detecting saliency on our multimodal images.

From Table 1 it is possible to notice that segmenting salient objects from images and videos [20] is able to generate more accurate saliency map on three modalities UVR, UVF, VIS images than other methods. This method performs better thanks to the variety of features used, followed by smoothing parameter and statistical operation. Spectral saliency detection [16] is better powerful in distinguishing salient region on the visible modality due to introducing reflectance attribute that characterized this modality. Itti model [14] shows a good score in IRR and UVF modalities. Context-based aware method [7] provides a good score in UVF modality, but receives a low-score in other modalities because it tends to favor the boundaries rather than interior region. ROI [4] and Signature-saliency [10] give a low-score according to the traditional background separation problem. While BMS [30] yields low score because it aims to extract the salient object rather than the salient pixel. SR also has a low-score because it does not interpret local saliency.

Acknowledgments. The authors thank the Chateau de Germolles managers for providing data and expertise and the COST Action TD1201 Colour and Space in Cultural Heritage (COSCH) (www.cosch.info) for supporting this case study. The authors also thank the PHC Toubkal/16/31: 34676YA program for the financial support.

References

1. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147**, 195–197 (1981)
2. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of attention and eye movements in real-world scenes: the role of global features in object search. Psychol. Rev. **113**, 766–786 (2006)
3. Anzid, H., Le Goic, G., Bekkarri, A., Mansouri, A., Mammass, D.: Improving point matching on multimodal images using distance and orientation automatic filtering. In: 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pp. 1–8 (2016)
4. Bharath, R., Nicholas, L.Z., Cheng, X.: Scalable scene understanding using saliency-guided object localization. 2013 10th IEEE International Conference on Control and Automation (ICCA), pp. 1503–1508 (2013)
5. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems, pp. 155–162 (2006)
6. Murabito, F., Spampinato, C., Palazzo, S., Pogorelov, K., Riegler, M.: Top-down saliency detection driven by visual classification. arXiv preprint arXiv:1709.05307 (2017)
7. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1915–1926 (2012)
8. Harel, J. Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2007)
9. He, S., Lau, R.W., Yang, Q.: Exemplar-driven top-down saliency detection via deep association. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5723–5732 (2016)
10. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: Zeng, Z., Li, Y., King, I. (eds.) Advances in Neural Networks – ISNN 2014. LNCS, vol. 8866, pp. 1–8. Springer, Cham (2007). https://doi.org/10.1007/978-3-319-12436-0_34
11. Hou, X., Harel, J., Koch, C.: Image signature: highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 194–201 (2012)
12. Hu, Y., Xie, X., Ma, W.-Y., Chia, L.-T., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004 Part II. LNCS, vol. 3332, pp. 993–1000. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30542-2_122
13. Han, J., Ngan, K.N., Li, M., Zhang, H.J.: Unsupervised extraction of visual attention objects in color images. Trans. Circ. Syst. Video Technol. **16**, 141–145 (2006)
14. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vis. Res. **40**, 1489–1506 (2000)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. PAMI **20**, 1254–1259 (1988)
16. Le Moan, S., Mansouri, A., Hardeberg, J.Y., Voisin, Y.: Saliency for spectral image analysis. J. Sel. Topics. Appl. Earth. Obs. Remote Sens. **6**, 2472–2479 (2012)
17. Sharma, P., Cheikh, F.A., Hardeberg, J.Y.: Saliency map for human gaze prediction in images. In: Sixteenth Color Imaging Conference, Portland, Oregon, USA (2008)

18. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: contrast based filtering for salient region detection. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–740 (2012)
19. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vis. Res. **45**, 2397–2416 (2005)
20. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010 Part V. LNCS, vol. 6315, pp. 366–379. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_27
21. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE 12th International Conference on Computer Vision, pp. 2106–2113 (2009)
22. Torralba, A.: Contextual priming for object detection. IJCV **53**, 169–191 (2003)
23. Torralba, A.: Modeling global scene factors in attention. JOSA **20**, 1407–1418 (2003)
24. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: CVPR, pp. 37–44 (2004)
25. Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: CVPR (2009)
26. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: ECCV (2012)
27. Yu, J.-G., Xia, G.S., Gao, C., Samal, A.: A computational model for object-based visual saliency: spreading attention along gestalt cues. IEEE Trans. Multimed. **18**, 273–286 (2016)
28. Chen, Z., Tu, Y., Wang, L.: An improved saliency detection algorithm based on Itti's model. Tehn. Vjesnik **21**, 1337–1344 (2014)
29. Hou, X., Zhang, L.: Saliency detection: a spectral approach. In: IEEE Conference on Computer Vision And Pattern Recognition (2007)
30. Zhang, L., Guo, C.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process. **19**, 185–198 (2010)
31. Chen, Z., Tu, Y., Eang, L.: An improved saliency detection algorithm based on Ittis model. Techn. Gaz. **21**, 1337–1344 (2014)