Fully Automatic Saliency-based Subjects Extraction in Digital Images

Luca Greco, Marco La Cascia and Francesco Lo Cascio
Dicgim, Università degli Studi di Palermo, Palermo, Italy

Keywords: Saliency Maps, Subjects Detection, Automatic Thumbnailing.

Abstract: In this paper we present a novel saliency-based technique for the automatic extraction of relevant subjects in digital images. We use enhanced saliency maps to determine the most relevant parts of the images and an image cropping technique on the map itself to extract one or more relevant subjects. The contribution of the paper is two-fold as we propose a technique to enhance the standard GBVS saliency map and a technique to extract the most salient parts of the image. The GBVS saliency map is enhanced by applying three filters particularly designed to optimize the performance for the task of relevant subjects extraction. The extraction of relevant subjects is demonstrated on a manually annotated dataset and results are encouraging. A variation of the same technique has also been used to extract the most significant region of an image. This region can then be used to obtain a thumbnail keeping most of the relevant information of the original image and discarding nonsignificant background. Experimental results are reported also in this case.

1 INTRODUCTION AND RELATED WORKS

The term saliency usually refers to visual quality or characteristic of interest for a human observer. It is often used also the term conspicuity, which literally indicates the visibility and what is most visible. In visual science a subject is important if it is characterized from what surrounds him, gaining greater visibility and attracting attention, resulting in a shift of the gaze. Therefore, the salient parts of a scene are those that evoke a strong visual response and polarize attention. Human visual attention is composed of two factors coming from two stimulus of different nature: the first one depends exclusively on the characteristics of the image, the second one is subjective in nature and is related to the subject's will. An objective stimulus has an bottom-up activation, due to the physical characteristics such as brightness, color, shape. In many situations the greatest contribution to the acquisition of information may be due to the top-down process, as the focus of the attention is influenced by the knowledge obtained by learning the probabilistic structure of the environment.

Saliency can be used in several tasks in Computer Vision. In this paper saliency maps are used as a basis to solve the subject extraction and thumbnailing problem in images.

The extraction of subjects in images consists of the identification of the regions where the most salient subjects of the image are located. Saliency is concentrated in the zones containing the subject, or subjects if there are more than one, and the rest of the image that can be considered as background, containing less informative pixels. The proposed algorithm is aimed to identify background and foreground areas and extract the subjects of the image.

The detection of salient subjects has many practical applications, such as cropping of images, extraction of dominant colors or other visual features of the objects of interest, search in image databases, etc.... There are several techniques to detect salient subjects. They can be divided into two main categories: the first category is based on the segmentation of the saliency map based on exhaustive search until it reaches a fixed fraction of saliency. Examples are the work of Marchesotti et al., (2009), which is based on a comparison of similar images stored on a database, the algorithm of Liu et al., (2007) based on finding the right weights for combining the various channels of the saliency map, and the algorithms of Cheng et al., (2011) and Wang et al., (2011) which are based on refinements of the saliency map. The second is based on the use of sliding windows on saliency map, which determine the probability of containing an object by determining areas with maximum saliency. These

Greco L., La Cascia M. and Lo Cascio F..
Fully Automatic Saliency-based Subjects Extraction in Digital Images.
DOI: 10.5220/0004530401290136
In *Proceedings of the 10th International Conference on Signal Processing and Multimedia Applications and 10th International Conference on Wireless Information Networks and Systems* (SIGMAP-2013), pages 129-136
ISBN: 978-989-8565-74-7

techniques are based on thresholds that identify the correct subjects and that are dependent on image data. Among these it is notable the work of Alexe et al., (2010) based on "'Objectness" and the work by Feng et al., (2011) based on the "superpixels".

Thumbnailing is widely used in many applications and consists of the creation of a smaller version of the image. This technique is used by visual search engines to organize and show images, by photo sharing sites and by all modern desktop operating systems. Even though this is an essential technique in many cases it is implemented as a simple image resizing. Image resizing is simple and fast but not always effective, since in resizing the image there is a loss of information that can make subjects not easily recognizable. Another possible implementation is based on cropping and resizing, like the one presented by Suh et al., (2003), which does not have the critical points of the simple resizing, but represents only a part of the image. It is also necessary to determine a criterion for cropping, which is a method to find a frame that contains what most distinguishes a photo. A natural criterion to measure the importance of a pixel according to Itti et al., (1998) is the use of a model of visual attention. For example in Suh et al., (2003) the authors proposes a compromise between the size of the crop and the internal value of the saliency. Another approach is to eliminate less important pixels trying to keep the high level semantic structures. Among these is the work of Samadani et al., (2007) that is based on the quick determination of natural previews. Avidan and Shamir (2007) present the "seam carving", but it run into problems if important content is spread all over the image. Simakov et al., (2008) presents another technique that overcomes, to a certain extent, the problems suffered by the seam carving. Other related work can be found in Liu and Gleicher, (2005:153-162) and Rother et al., (2005).

2 SALIENCY

The analysis of an image is characterized by what one would expect to find or what you want to identify. By applying an attention based strategy it is possible to reduce the time of search or analysis. In summary attention is composed of two separate processes, as claimed by Itti and Koch (1998):

- A bottom-up process, which is based exclusively on a physical stimulus. This process is simple and fast but lacks from any high level considerations.
- A top-down process, which depends on observer will and that is based on a semantic analysis of the image. This process has a higher level of abstraction but it is slower. This search is more accurate having a high degree of environmental knowledge and examing the consistency of the context.

In this paper we will deal with bottom-up stimulus and in particular we use the saliency map model presented in Harel et al., (2006).

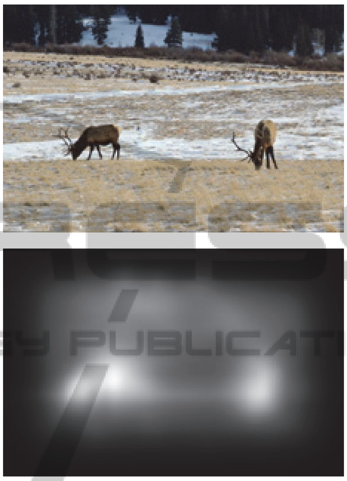

Figure 1: Original image and GBVS saliency map.

2.1 Graph-based Visual Saliency

The GBVS is a model of bottom-up visual saliency, which is composed of two stages: formation of the activation maps of certain features such as color, intensity, brightness, and feature normalization to emphasize the saliency and to allow combination with other maps.

First stage consist in extracting feature maps and creating activation maps, constructing a fully-connected graph with a node for each pixel and directed edges weighted using pixels' dissimilarity in feature map and their distance. The weights are normalized to 1 and used as the transition probability of a Markov chain, then the activation map is calculated by the equilibrium distribution of the resulting chain. Second stage is performed constructing another graph using the calculated activation map and the distances. Again, this graph is used as a Markov chain and the normalized map is derived by the equilibrium distribution.

2.2 Saliency Map

The GBVS saliency map is represented by a matrix representing the saliency of each pixel. The map is normalized, containing values between zero and one: the closer the saliency is to one, more salient is the corresponding pixel. An example of saliency map is shown in Figure 1 where white pixels correspond to a saliency value equal to one, and black for a saliency value equal to zero. Therefore in correspondence of the two deers there is a higher concentration of the salience while in the areas of the background, i.e. the trees in the distance and the terrain, the concentration of saliency is lower and almost uniform.

2.3 Noise in the Map

In some cases the presence of noisy and non-uniform background in the image affect the correct determination of the saliency map. Also in images where many subjects are concentrated in a small area, the saliency map tends to consider the whole area salient without distinguishing between the various subject. Moreover, if a subject has characteristics (shape, color, orientation, etc.) similar to the background it is considered not salient and not detected as a subject of the image.

3 PROPOSED ALGORITHM

The proposed algorithm is designed to solve two different problems:

- The detection and extraction of interesting subjects. This includes the subjects detection in an image and determining for each of these subjects a bounding box.
- Determination of a significant thumbnail that is the determination of a small preview picture of the original image.

The algorithm is based on a saliency map to determine in the first case the areas where interesting subjects are present and to determine in the second case the foreground and the background.

In both cases, the structure of the algorithm remains the same changing only the stopping criterion.

3.1 Computation and Refining of Saliency Maps

The saliency map used is obtained by GBVS algorithm, so it is purely dominated by objective stimulus without taking into account the global image. For the task of the proposed method the interesting part of the map is the one associated to subjects, so eliminating the part associated to background can lead to a better result. If in the image there are one or more subjects a more useful map is composed of saliency spots corresponding to the subjects.

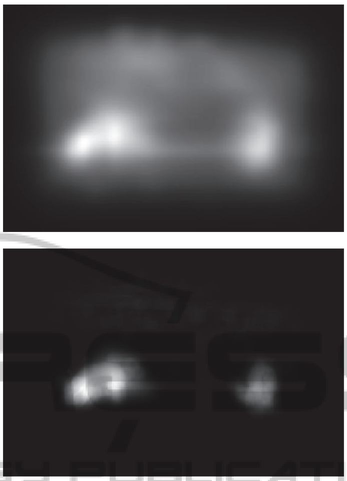

Figure 2: Initial map and refined version.

The application of three filters, each with a specific goal, can lead to such a map. The first filter accentuates the saliency in correspondence of image contours, to make the subject cropping more precise and decrease the possibility of an incomplete cut. In fact, if the subject has not a compact shape it may occur that the contour details are ignored in the saliency map. To prevent this, a binary map describing the pixels that belong to the contours is used. So in the saliency map the pixels corresponding to ones in the contour map are incremented by a factor of 0.3. Contours are extracted using a Sobel filter.

The second filter deals with transforming the saliency map, through a nonlinear transformation, increasing the difference between background pixels and those which belong to subject. After this transformation less informative pixels are truncated using a fixed threshold. The nonlinear transformation is:

$$y = x^4. \tag{1}$$

Finally, the third filter is a Gaussian filter that has the result of smoothing the refined saliency map. Figure 2 shows the result of these filters applied to the saliency map of the two deers.

3.2 Image Cropping

Based on the refined saliency map we define a cropping algorithm based on a greedy search. A greedy algorithm attempts to construct a pseudo-optimal solution from a partial initial solution and searching until a stop criterion. The extensions of the solution do not consider all possible solutions because this could be very expensive. Only some solutions are considered in paricular those that are closer to the partial solution, proceeding to the attainment of the solution pseudo-optimal for small steps. «

These extensions of a partial solution that could be called "local" are somewhat smaller and possible extensions are relatively few. Among the various local extensions the algorithm proceeds to choose the most "greedy", or rather the most convenient. So, the more promising extension that is configured as a local optimal solution is selected as a new partial solution.

The cropping algorithm based on saliency is initialized with a partial solution obtained by cutting a box centered at the peak of the salience of fixed size (5x5 pixels is the proposed initial box). Then we proceed to increase the dimension of the box obtained at each iteration of partial solutions considering the most valid solutions the one with a higher percentage of salience.

Partial solutions are selected in such a way to obtain at each iteration a pane that possesses an increasing saliency and possible solutions are computed changing two factors:

1. The center of the frame. The center of the subject does not correspond with the peak saliency is then evaluated the increase of saliency by modifying the center of the box along the diagonal directions
2. The increase of the box. In most cases the optimal cropping is not a square, thus evaluating width or length direction.

3.3 Stopping Criterion

The stopping criterion should determine if the partial frame obtained contains sufficient saliency area of the subject and therefore if it defines the final rectangle. At each iteration the difference of the saliency of last frame and the new proposed frame is evaluated and the search algorithm is stopped when the increase of salience is under a fixed threshold close to zero. Another stopping criterion is evaluating the sum of internal saliency of the frame and stop if it exceeds a threshold.

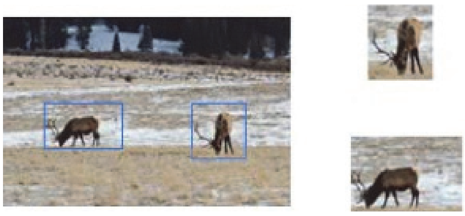

Figure 3: Query image and subjects detected.

4 SUBJECT DETECTION

The proposed method can be used to identify subjects in images, if presents. For this problem, images can be divided in two categories, like in Huang et al., (2010):

Salient Images: containing one or more subjects easily identified, placed in a relatively uniform background.

Cluttered Images: difficult to distinguish subjects, with disordered background and high presence of noise and details. Extracted saliency map can be influenced by uninformative part of background or by details.

Subject extraction problem can be stated as follows: given an image, the aim is to find a rectangle, containing a subset of image where the subject is visible. In this problem, using the saliency value for evaluating the degree of informativeness of different areas of the image, the crop rectangle must meet two conditions: have a small size and contain most of the subject. These two conditions are usually in conflict each other, therefore, the goal is to find the pseudo-optimal rectangle with a trade-off of the constraints.

4.1 Subjects Extraction

An image can represent more than one salient subject, but the proposed algorithm can extract a single entity starting by a peak of salience. Therefore, it is possible to reuse the same algorithm to extract other subjects in the picture by ignoring the part selected for the first subject and considering the remaining saliency map to determine the presence of other subjects in the picture.

The map is then processed again by the algorithm that will identify another peak salience and extracts another frame. This technique leads to the extraction of all salient areas in the image related to the subjects. Finally, the operation is repeated until the percent of the salience remained after the last cut is less than 5% of the initial saliency. If we apply this algorithm to the image of two deer we obtain two frames containing the two subjects. In fact, in Figure 3 we can see the previews of the subjects they represent completely, giving two frames that isolate the subjects from the background.

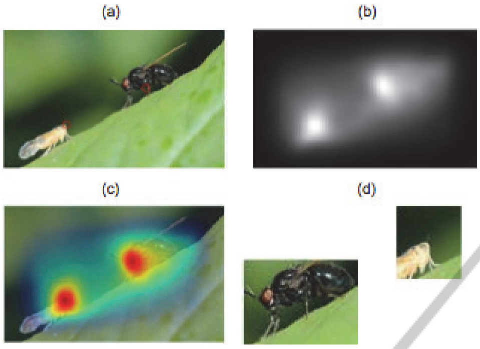

Figure 4: Not precise saliency map lead to a not correct extraction.

4.2 Incomplete Preview

In some cases, the object composition and structure is not accurately identified by the saliency map. This is due to the nature of the saliency map that is exclusively linked to the features and does not take into account the semantics of the image. In fact, the saliency map can be erroneously determined and parts of the main subject may not be detected because they were considered not salient. It may also occur that parts of the background may be considered an integral part of a subject. These two phenomena lead to a lack of precision of the saliency map and are reflected in the results of cropping, making the previews inaccurate. In some cases previews of images that include parts of the background where computed. Although the subject remains recognizable, the frame is not accurate. In other cases images contain only part of the subject that can be interpreted as the particular most dissimilar from the rest of the image. In these cases the saliency has a sharp peak and the rest of the subject is consider not information. An example may illustrate these cases (Figure 4). It can be observed that the salience in an area of the background in proximity of the white insect is not distributed with precision on the subject. This moves the crop from the subject to the background by perturbing the correctness of the final frame. The preview that appears in fact not entirely contains the subject but only a part of the latter and part of the background.

4.3 Results Metrics

To evaluate the results of subjects detection two factors are considered: the number of subjects revealed and the accuracy of the crop. In fact, it is necessary to first evaluate how many subjects were identified by the algorithm on the total number present in the image and then assessing if the crop computed contains entirely or only a part of the subject.

We used two different datasets. A dataset of 20 images with salient subjects have been identified for the assessment of the estimation the number of subjects. Approximately 80% of the total number of interesting subjects are correctly estimated. The method show the tendency to not overestimate the number of interesting subjects. When the detection is wrong usually the estimated number of subjects is less then real. On a total of 20 images only in one case the number of salient subjects estimated was larger then real one.

For the evaluation of the accuracy of the crop some pictures of the dataset presented in Wang et al., (2012) have been used. The images of the dataset contain a main subject and a frame containing it has been annotated by hand. Figure 5 show an example of this annotation.

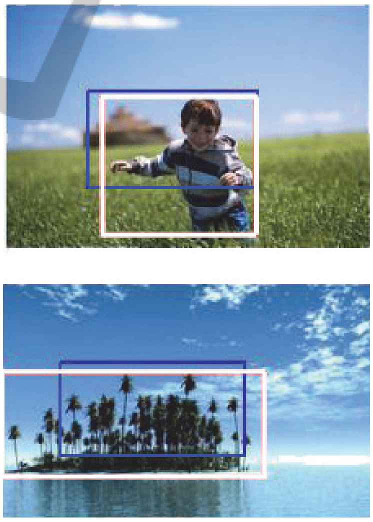

Figure 5: The blue box is the result of the proposed algorithm, the white box is the manual crop of the subject.

Therefore, it is possible to determine how the frame identified by the algorithm, Ralg, is similar to the frame noted, Rnot, using expressions:

$$\text{Precision} = \text{Area}\,(R_{alg} \cap R_{not}) \,/\, \text{Area}(R_{alg})$$

$$\text{Recall} = \text{Area}\,(R_{alg} \cap R_{not}) \,/\, \text{Area}(R_{not})$$

where the precision indicates percentage of the area of Ralg that is contained in Rnot and the recall indicates the tendency of Ralg to contain the frame annotated. In the experiments precision has an average value of 0.68 and recall of 0.75 .

The measures are discordant with each other and can be enclosed in a single greatness, called F-measure, and defined mathematically:

$$F_\alpha = \frac{(1+\alpha) * Precision * Recall}{\alpha * Precision + Recall} \qquad (2)$$

where for our purposes α is set to the value of 0.5. In a dataset of 40 image has been achieved on average $F_{0.5} = 0.66$.

All the dataset used for the experiments will be available online soon.

5 THUMBNAILING

A modified version of the proposed algorithm can also be used for image thumbnailing. In subject extraction the focus was on finding spots in the refined saliency map, so the stopping criterion is aimed to find the rectangle that contain a single spot. The process is then repeated until all subjects are extracted. For thumbnailing the goal is to find a single rectangle containing the most recognizable part of the original image. It is then sufficient to resize this selected part of the original image to obtain a smart preview.

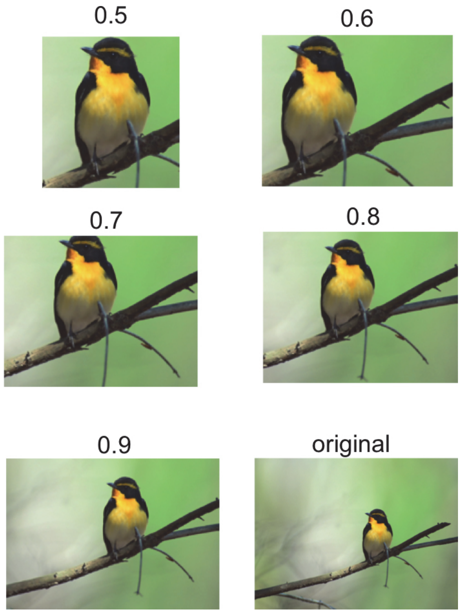

Figure 6: Not constrained crop.

The stopping criterion for this problem is based on thresholding of the ratio of the sum of saliency in the proposed crop and the sum of saliency over the entire map. Changing the threshold from low to high values the proposed cropping rectangle raise in dimension and contain most salient part of the initial image. Figure 6 shows the results for different threshold values. The higher is the threshold the larger is the dimension of the selected thumbnail, starting from a close view of the subject.

Using the proposed algorithm, the result of the crop is a rectangle which width and height proportion are not fixed. If there is a need for a particular aspect ratio of the preview image, for example in an image browsing application, the greedy policy of search can be changed to obtain a fixed aspect ratio. For the example in Figure 7 and in the following section a modified version is used that perform only square crop.

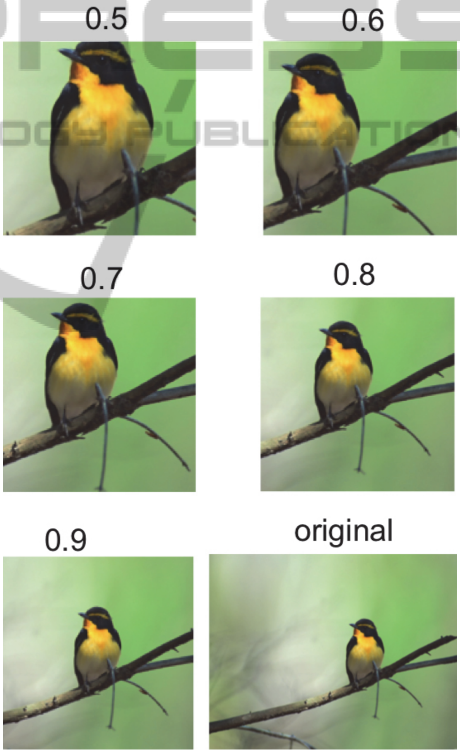

Figure 7: Square crop.

Visually analyzing the result of thumbnailing we found that a good value for the threshold is 0.5.

5.1 Yahoo! Thumbnails Comparison

The results of the proposed method were compared with thumbnails extraction of Yahoo! (www.yahoo.com) that propose square previews as result of an image search. Figure 8 shows the difference between the different thumbnails for the images obtained using the keyword "Animals". In column (a) are shown the original images, in column (b) the previews extracted from Yahoo!, in column (c) our results.

The proposed algorithm generally focuses on salient part producing a reduced part of the original image while Yahoo! seems to simply crop the image.

Generally, this representation is useful if the previews are used for searching for details and content of pictures or known image. High presence of background in the preview reduce the information shown. In personal photo collection, for example, having a salient and zoomed part of an image can lead to a faster visual retrieval of the image.

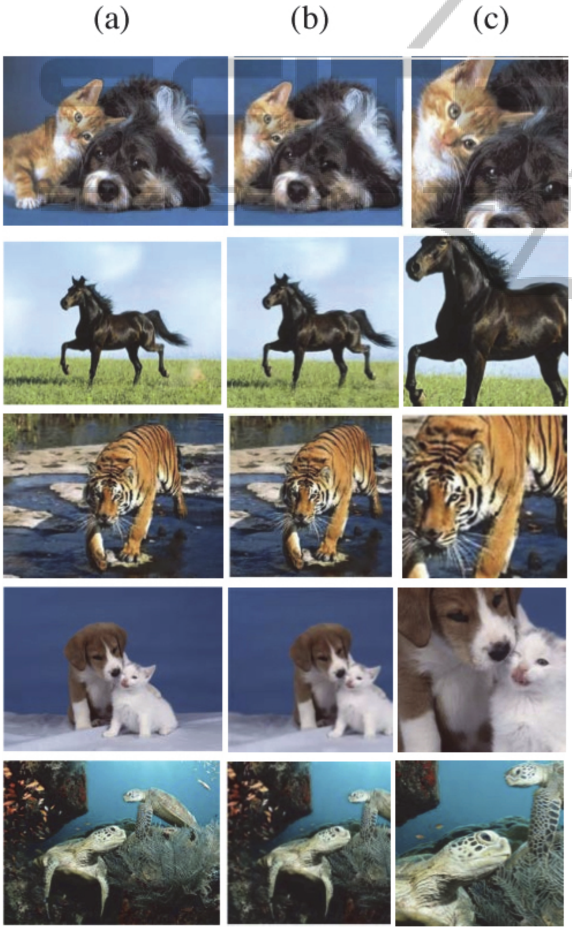

Figure 8: Comparison of yahoo thumbnails and the result of the proposed method.

6 CONCLUSIONS

We presented a simple technique to automatically extract relevant information from digital images using a bottom-up visual saliency model.

The technique has been applied to detect the most relevant subjects in the foreground also in the case of cluttered background. Performance were evaluated on small annotated image dataset we prepared for our experiments and that will be available for download and could be used for future comparison.

A variation of the same technique has also been demonstrated on the problem of smart thumbnailing that is the creation of thumbnails keeping most of the information of the original image using a much smaller number of pixels. A qualitative comparison with Yahoo! thumbnails has shown also in this case the goodness of our approach.

In both cases we observed that filtering the saliency map significantly increases the performance.

Finally, even though in many cases a top-down model-based approach might be preferable we have shown that, in many cases, a simple data driven approach could be satisfactory.

REFERENCES

Harel, J., Koch, C. and Perona, P., 2006. Graph-Based Visual Saliency. *Proceedings of Neural Information Processing Systems (NIPS).*

Marchesotti, L., Cifarelli, C. and Csurka, G., 2009. A framework for visual saliency detection with applications to image thumbnailing. *Xerox Research Centre Europe (XRCE),* France

Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W., 2003. Automatic Thumbnail Cropping and its Effectiveness. *Proceedings of the 16th annual ACM symposium on User interface software and technology*

Liu, L., Chen, R., Wolf, L., Cohen-Or, D., 2010. Optimizing Photo Composition. *Computer Graphics Forum*, vol. 29, n. 2 pp. 469-478, Wiley Online Library

Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W., 2005. Auto Cropping for Digital Photographs., *IEEE International Conference on Multimedia and Expo, (ICME)*.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X. and Shum, H.-Y., 2007. Learning to detect a salient object. *IEEE Trans. Pattern Anal*. Mach.

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X. and Hu, S.-M., 2011. Global contrast based salient region detection*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Wang, L., Xue, J., Zheng, N. and Hua, G., 2011. Automatic salient object extraction with contextual cue. *IEEE Conference on Computer Vision (ICCV).*

Alexe, B., Deselaers, T. and Ferrari, V., 2010. What is an object*? IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Feng, J.,Wei, Y., Tao, L., Zhang, C. and Sun, J., 2011. Salient object detection by composition. *IEEE Conference on Computer Vision (ICCV)*.

Huang, J., Yang, X., Zhang, R. and Fang, X., 2010. Re-Ranking Image Search Results by Multiscale Visual Saliency Model. *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)* .

Itti, L., Koch, C. and Niebur, E., 1998, A model of saliency-based visual attention for rapid scene analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 20 n. 11 pp. 1254-1259.

Samadani, R., Lim, S. H. and Tretter, D., 2007. Representative image thumbnails for good browsing. *IEEE International Conference on Image Processing (ICIP)*.

Avidan, S. and Shamir, A., 2007. Seam carving for content-aware image resizing. *ACM Transaction on Graphics*.

Simakov, D., Caspi, Y., Shechtman, E. and Irani, M., 2008. Summarizing visual data using bidirectional similarity. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Liu, F. and Gleicher, M., 2005. Automatic image retargeting with fisheye-view warping. *ACM symposium on User interface software and technology (UIST)*.

Rother, C., Kumar, S., Kolmogorov, V. and Blake, A., 2005. Digital tapestry. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, P., Wang, J., Zeng, G., Feng, J., Zha, H. and Li, S., 2012. Salient Object Detection for Searched Web Images via Global Saliency *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*