Saliency Based Image Cropping

Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola

Dipartimento di Ingegneria Chimica, Gestionale, Informatica, Meccanica,
Università degli studi di Palermo, Italy
{edoardo.ardizzone,alessandro.bruno15,giuseppe.mazzola}@unipa.it

Abstract. Image cropping is a technique that is used to select the most relevant areas of an image, discarding the useless ones. Handmade selection, especially in case of large photo collections, is a time consuming task. Automatic image cropping techniques may help users, suggesting to them which part of the image is the most relevant, according to specific criteria. We suppose that the most visually salient areas of a photo are also the most relevant ones to the users. In this paper we present an extended version of our previously proposed method, to extract the saliency map of an image, which is based on the analysis of the distribution of the interest points of the image. Three different interest point extraction algorithms are evaluated within an automatic image cropping system, to study the effectiveness of the related saliency maps for this task. We furthermore compared our results with two state of the art saliency detection techniques. Tests have been conducted onto an online available dataset, made of 5000 images which have been manually labeled by 9 users.

Keywords: Image Cropping, Visual Saliency, Visual Perception, Saliency Map.

1 Introduction

The development of network and computer hardware technology, and mobile media devices (smartphones, digital cameras and PDA) has had a large impact in people's everyday life. Today many people use these devices to take a lot of photos, and share them by social networks. Image cropping is a technique that is used to resize an image by selecting its most relevant areas, discarding its useless or redundant parts. People may wish to remove portions of photo background to emphasize the subject (fig.1), to fit an image to fill a frame, to create thumbnails for browsing purposes, or to select the most important (to the observer) parts of the image, to improve photo-composition. In standard photo-editing applications, designers manually crop an area around the important content of the image. Some commercial products allow users to manually crop images to generate thumbnails. Handmade cropping, in case of large photo collections, is an onerous and time-consuming task. Automatic cropping methods suggest to the users which are the most important parts of the image. What does it mean for "the most important parts of a photo" in this context? In our work we suppose that the most salient regions of the image are considered as the most relevant parts of a photo. The aim of visual saliency detection methods is to build a saliency map that replicates the human visual

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 773–782, 2013.

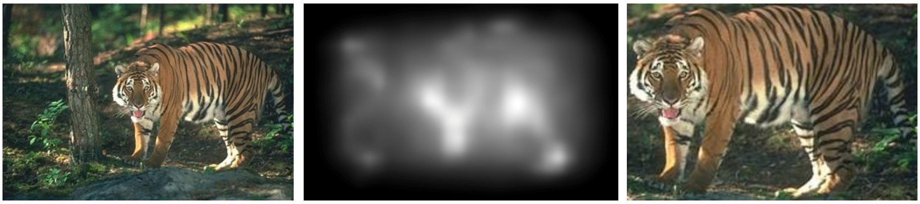


Fig. 1. Input Image (left), saliency map (center) and image crop (right)

system (HVS) behavior in the visual attention process. From a visual perception viewpoint, many images are made of salient regions surrounded by unnecessary background areas. In this paper, we present an extended version of our previous work on visual saliency[1] and we study the effectiveness of saliency maps in image cropping methods. We compare the results obtained by our method with those obtained with two state of the art techinques, within a common automatic image cropping system. We evaluate results using a free available dataset made of manually cropped images (more details about the dataset are given in section 5). The paper is organized as follows: in section 2 we discuss some State of the art methods about Image Cropping; in section 3 we discuss our saliency detection algorithm and the two reference methods; in section 4 we describe the Saliency-Based Image Cropping system; in section 5 we show and discuss the experimental results; section 6 contains conclusive considerations.

2 State of the Art

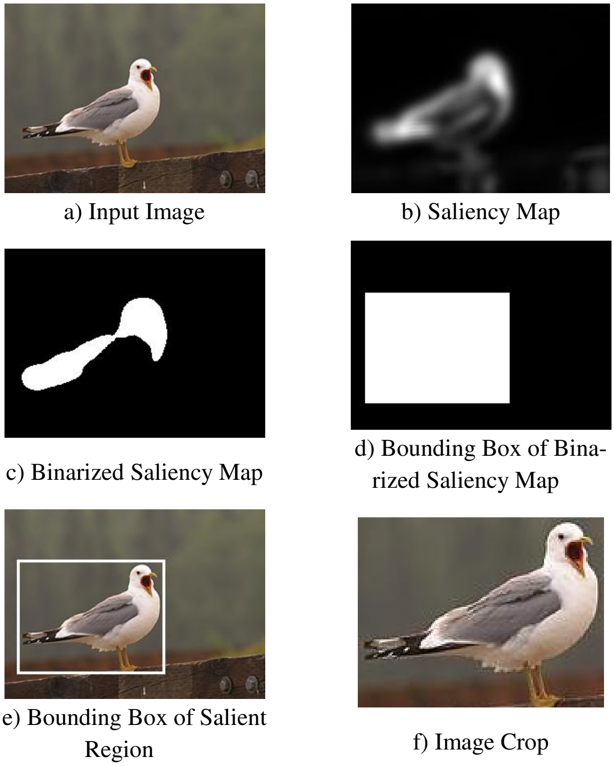
In the last years, there has been much interest in automated cropping methods, especially using visual attention information. Suh et al. [2] used both saliency maps and automatic face detection to evaluate candidate cropped regions to determine best crop. Ma et al. [3] first segmented the images, then ROI are selected according to the image entropy, the size and the closeness to the center of the image. Zhang et al. [4] formulated automatic cropping as an optimization problem, which consists of three sub models (composition, conservative, and penalty), and employed a particle swarm optimization (PSO) to obtain the optimal solution by maximizing the objective function. Santella et al. [5] employed an eye tracking system to identify the important content of a photo, for automatic snapshot recomposition, adaptive documents, and thumbnailing. Stentiford [6] proposed a method for automatically cropping photos and camera zooming based upon a new visual attention model. Ciocca et al. [7] proposed a self-adaptive image cropping algorithm where the processing steps are driven by the classification of the images into semantic classes, exploiting both visual and semantic information. In She et al. [8] the authors first classified photos in five categories, and then build a dictionary for each category, extracting the visual saliency maps of these photos. Some more recent works on image cropping emphasize the pleasantness of resulting cropped photos, taking into account photographic composition rules (as the rule of thirds, the rule of filling the frame and leading the lines) or quality scores. In Luo [9] the author performed image cropping method based on subject detection algorithm, using a belief map that probabilistically indicates the subject content. Nishiyama et al. [10] build a quality classifier using a dataset of photos into which users manually insert quality scores to photos. They finally used the classifier to find the cropped region with the highest quality score. Cheng et al. [11] proposed a photo quality evaluation metric for automatic professional view finding, exploiting professional photographers' knowledge and composition rules. Bhattacharya et al. [12] presented an interactive application that helps users to improve the visual aesthetics of their digital photographs using spatial re-composition. Liu et al. [13] applied the rule of thirds, the diagonal dominance, visual balance and sizes of salient regions for equally evaluation. Ahn et al. [14] used crowdsourcing techniques to collect many crops for a set of 68 photographs. They analyze the crops with respect to the composition guidelines recommended in photography and art literature. In McManus et al. [15] the authors analyze the psychic aspects of Image Cropping and the influence of color, semantics and expertise of the users on the resulting crops.

3 Visual Saliency

Visual Saliency deals with identifying the most important regions of an image from a perceptual point of view (Frintrop et al. [16]). In the first three seconds a human observer fixates some particular points inside an image and tends to group them into visual significant areas (Judd et al. [17]). In [18] Achanta et al. exploit features of color and luminance to detect salient objects into the image. In this paper we compared five saliency maps to evaluate the effectiveness these methods, for image cropping applications: Itti et al. [19], Harel et al. [20], and three variations of our previous work [1]. All the analyzed methods are bottom-up, stimulus-driven and unsupervised:

- Itti-Koch model [19] is based on a multi-scale analysis of the image. Multi-scale image features are combined into a single topographical saliency map. A dynamical neural network selects attended locations in descending order of saliency.
- Harel [20] saliency approach is based on a biologically plausible model, and it consists of two steps: activation maps on certain feature channels and normalization, which highlights conspicuity.
- In our method [1] we analyze the distribution of the keypoints onto the image, with different scales of observation. In particular SIFT-point Density Maps (SDM) are built to study the relationship between the keypoints extracted by the SIFT algorithm [21] and real human fixation points. We furthermore extended our previous work by considering also other two types of keypoint extraction algorithms: Harris Corner [22] and SURF[23]. More particularly, our previous method [1] analyzes the distribution of the SIFT interest points along the image, and build the SIFT Density Maps (SDM). The most salient areas are those that maximizes the difference between the SDM value and the most frequent value of the SDM (which we suppose related to the background).

In the remainder of the paper we will refer to them with ITTI, GBVS, SIFT, SURF and HARRIS, respectively.



a) Input Image

b) Saliency Map

c) Binarized Saliency Map

d) Bounding Box of Binarized Saliency Map

e) Bounding Box of Salient Region

f) Image Crop

Fig. 2. The steps of the Image Cropping Method pipeline

4 Saliency Based Image Cropping

The aim of the present work is to evaluate the effectiveness of saliency maps when used to support automatic cropping. Our system is subdivided into (see fig.2): Saliency Map Extraction, Saliency Map Binarization (Thresholding), Bounding Box Extraction, Photo Cropping, Evaluation. Given an image, we compute the five saliency maps described in section 3 (ITTI, GBVS, SIFT, SURF, HARRIS). Each saliency map is then binarized using different threshold values (see section 5) and then the bounding box of all the pixels, which values are above the threshold, is selected and used to crop the photo (fig.2). Results are evaluated (in terms of precision, recall and F-measure) comparing the resulting crops with handmade selected crops, from the dataset created by Liu et al. [24], that will be further described in the next section. We selected ITTI and GBVS as reference methods to compare, as they are the most cited methods in literature, and they obtain very good results in any application in which they are used. The center-weighted Gaussian estimator is also a solution but, to our knowledge, both ITTI and GBVS typically achieve much better results. Moreover, the photographic "rule of the thirds" suggests that the subject of the photo should not be placed in center of the image, but along the intersections of a 3x3 grid, superimposed to the image. Therefore, the "center" method will not work in these cases.

5 Experimental Results

Our experiments has been conducted onto a freely available dataset [23] which consists of 5000 images labeled by 9 users, who have been invited to select the most salient object of the scene represented in the image, by drawing a rectangle. For our purposes, we compute the "average crop" of the handmade crops of this dataset considering, for each image, only the pixels that have been selected at least by N

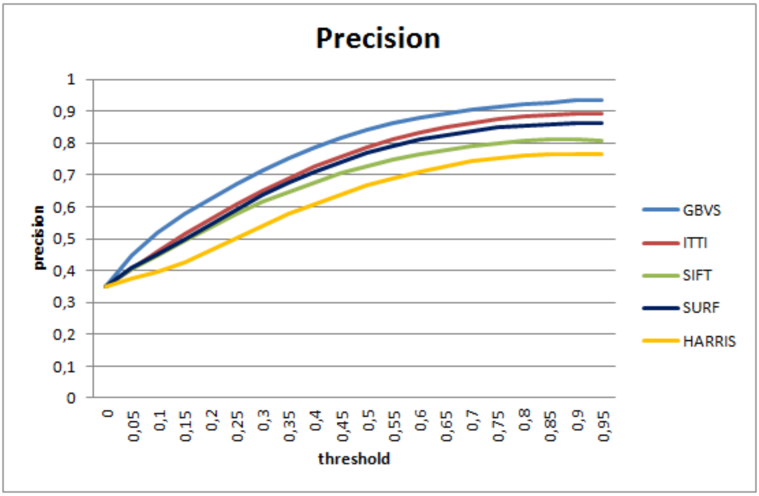


Fig. 3. Experimental Results: Precision

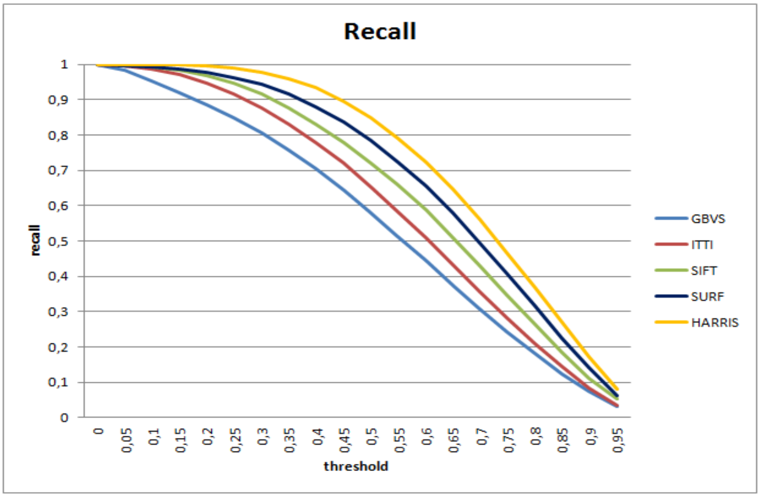


Fig. 4. Experimental Results: Recall

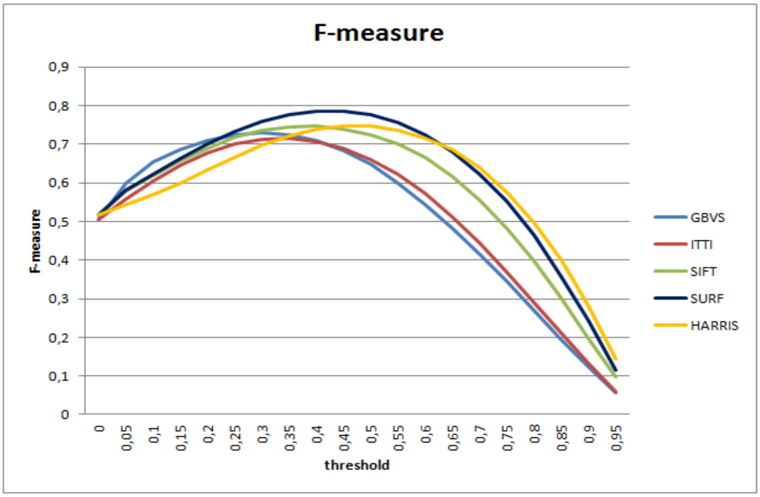


Fig. 5. Experimental Results: F-measure

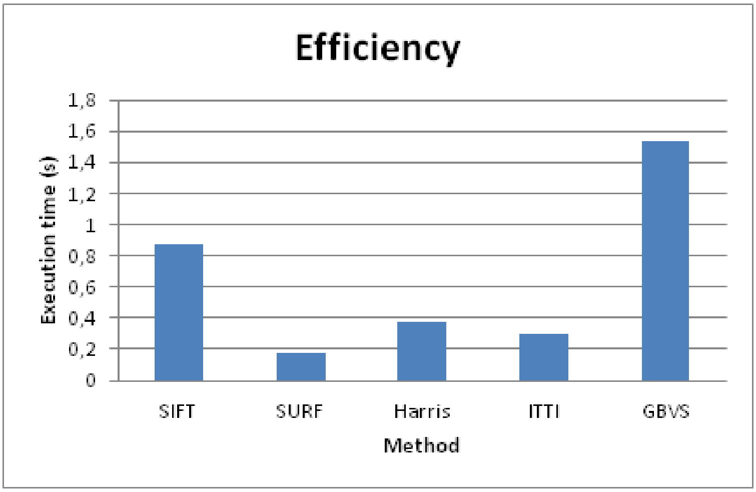
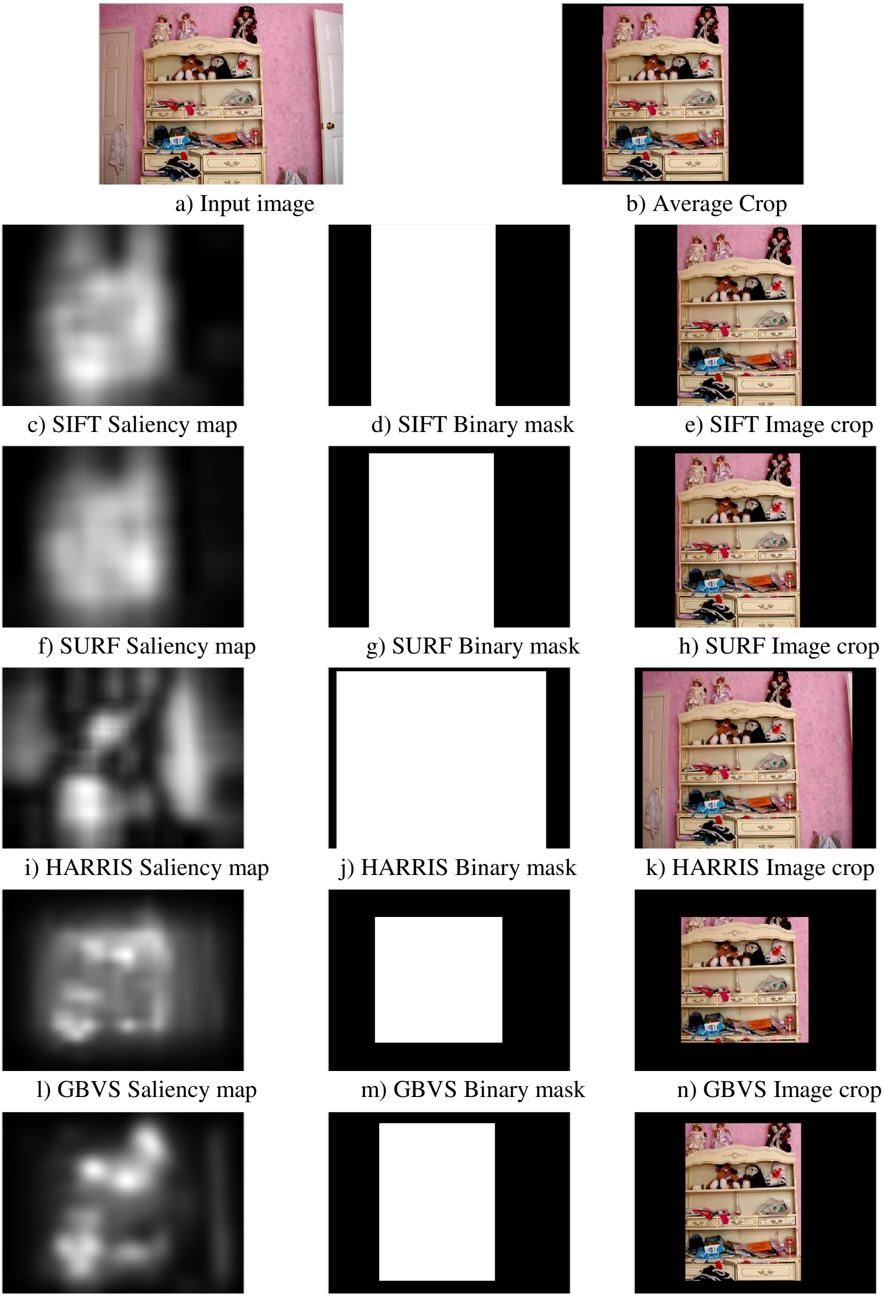


Fig. 6. Experimental Results: Efficiency

(in our experiments N=5) of the 9 users, i.e. the majority of the users. For each of 5000 images of the dataset, we computed the five saliency maps (ITTI, GBVS, SIFT SURF, HARRIS) and we binarize them by using different threshold values *T* (from 0 to 0.95 with step 0.05). An interesting study could be to analyze the relationship between the threshold and the content of the images. This study could be done only after having labelled the images in the dataset, in terms of some pre-established classes. But this is not the focus of our paper. The accuracy of the results is measured



a) Input image

b) Average Crop

c) SIFT Saliency map

d) SIFT Binary mask

e) SIFT Image crop

f) SURF Saliency map

g) SURF Binary mask

h) SURF Image crop

i) HARRIS Saliency map

j) HARRIS Binary mask

k) HARRIS Image crop

l) GBVS Saliency map

m) GBVS Binary mask

n) GBVS Image crop

o) ITTI Saliency map

p) ITTI Binary mask

q) ITTI Image crop

Fig. 7. A visual example of image crops with the different saliency maps. Saliency maps are binarized with threshold = 0.35, which is a good tradeoff for all the methods (see fig. 6)

comparing the binary mask C_S of the crops (fig. 2.d), for a given threshold, with the binary mask C_A of the "average crop", in terms of recall, precision and F-measure:

$$P = \frac{n(C_S \cap C_A)}{n(C_S)} \tag{1}$$

$$R = \frac{n(C_S \cap C_A)}{n(C_A)} \tag{2}$$

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{3}$$

where:

- R is the recall, the ratio of the number of pixels in the intersection of the saliency crop C_s and the average crop C_A, and the number of pixels in C_A;
- P is the precision, the ratio of the number of pixels in the intersection of the saliency crop C_s and the average crop $C_{A,}$ and the number of pixels in C_s;
- F_1 is the F-measure.

Figures 3,4,5 show Precision, Recall and F-measure, averaged on the 5000 images in the dataset. Note that the "average crop" is independent from the threshold values, while the automatic cropped areas vary with these values. The first important result is that the all the methods achieve very good results in terms of precision (fig. 3), for most of the threshold values. GBVS and ITTI are the best ones, while keypoint-based methods have slightly worse results. In terms of recall (fig. 4), SIFT, SURF and HARRIS achieve better results than GBVS and ITTI. It means that GBVS and ITTI return smaller crops which, however, includes fewer pixels of the related average crops. The last three methods return larger crops, which include more pixels of the average crop areas, but a little bit more false positives. In terms of F-measure (fig. 5), keypoint-based methods achieve better results than GBVS and ITTI, and SURF above all. About efficiency, fig. 6 shows the execution times of the compared methods, averaged for the 5000 images of the dataset. The fastest method is SURF, while the slowest is GBVS. For the keypoint-based methods, most of the time is spent to extract keypoints, in fact the time to build the map is one of two orders of magnitude smaller than the time to extract points. Moreover, with respect to our previous version[1], we improved in terms of efficiency our implementation of the map building algorithm. In our previous work, the saliency map was built by shifting a window along the pixels of the image and counting the number of keypoints it includes, to study the distribution of the keypoints along the image. The newest version of our method focuses on the keypoints, that is, for each keypoint we update simultaneously the values of all the windows that will include it, drastically reducing the execution time. Finally, fig.7 shows some visual results obtained with the different saliency maps.

6 Conclusions

In this paper we present an extended version of our previous work on visual saliency and we evaluated its effectiveness when used to support automatic image cropping.

Saliency-based crops have been compared to handmade crops, within a standard database. We improved the algorithm implementation, as briefly described in section 5, in terms of efficiency, as our previous version was slower than the reference methods, while the new one is comparable or faster.

Note that a user took typically 10-20 seconds to select the part to crop and to draw a rectangle onto an image, while (bottom-up) saliency detection methods typically aim to reproduce the behavior of the Human Visual System in the very first instants when observing an image. Therefore there is a time "gap" between the moment in which the user recognizes a salient part of an image and the moment in which he manually crops that area. After the first 3 seconds users are guided, when analyzing a scene, also by high-level mental processes (recognizing objects, context, faces), therefore there can be a difference between what is visually salient and what is representative of an image. Results showed that saliency-based approaches are very suitable for automatic image cropping applications. We suppose that results could be further improved if different cropping strategies are adopted for different categories of image, as images can represent scene with any type of visual and semantic content.

Regarding the "multiple objects" question, our work is inspired by the assumption that, according to the photocomposition rules, a "high quality" photo must have only one and distinct subject. This is only a suggestion and not a strict constraint, and it is not true for some categories of photo (e.g. panoramas). We think that saliency based methods do not work as well in case of multiple objects. Probably they will include in the same crop all the salient objects (if they are close enough), or they select only one of the salient object, if they are far. Therefore it strongly depends on the reciprocal distance between the salient objects in the scene. Further experiments are needed to face this specific problem.

In fact we observed that saliency based automatic cropping methods, as expected, give worse results when the background area is very composite, or whenever it is not easy to detect a single foreground object into the scene. In those cases, for automatic cropping an image, saliency detection methods could be supported by segmentation algorithms as a preprocessing step. The method can be further improved when applied to other tasks, e.g. face or object detection, if combined with other types of information, e.g. from face detector or object classifiers. But this is not the focus of the paper, which intend to be general purpose. We intend to further study this problem in our future works.

References

1. Ardizzone, E., Bruno, A., Mazzola, G.: Visual saliency by keypoints distribution analysis. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part I. LNCS, vol. 6978, pp. 691–699. Springer, Heidelberg (2011)
2. Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic Thumbnail Cropping and its Effectiveness. In: Proc. of the 16th ACM Symposium on User Interface Software and Technology, pp. 95–104 (2003)
3. Ma, M., Guo, J.K.: Automatic Image Cropping for Mobile Devices with Built-in Camera. In: Proc. of the Consumer Communication & Networking Conf., pp. 710–711 (January 2004)

4. Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W.: Auto Cropping for Digital Photographs. In: IEEE International Conference on Multimedia and Expo (2005)
5. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.F.: Gaze-based interaction for semi-automatic photo cropping. In: CHI 2006, pp. 771–780 (2006)
6. Stentiford, F.: Attention Based Auto Image Cropping. In: ICVS Workshop on Computational Attention & Applications (2007)
7. Ciocca, G., Cusano, C., Gasparini, F., Schettini, R.: Self-Adaptive Image Cropping for Small Displays. IEEE Trans. on Cons. Electronics 53(4), 1622–1627 (2007)
8. She, J., Wang, D., Song, M.: Automatic image cropping using sparse coding. In: 2011 First Asian Conference on Pattern Recognition (ACPR), pp. 490–494 (2011)
9. Luo, J.: Subject Content-Based Intelligent Cropping of Digital Photos. In: IEEE International Conference on Multimedia and Expo, pp. 2218–2221 (2007)
10. Nishiyama, M., Okabe, T., Sato, Y., Sato, I.: Sensation-based photo cropping. In: Proceedings of the 17th International Conference on Multimedia 2009, pp. 669–672. ACM, Vancouver (2009)
11. Cheng, B., Ni, B., Yan, S., Tian, Q.: Learning to photograph. In: Proceedings of the International Conference on Multimedia, Ser. MM 2010, pp. 291–300. ACM (2010)
12. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: Proceedings of the International Conference on Multimedia (MM 2010), pp. 271–280. ACM, New York (2010)
13. Liu, L., Chen, R., Wolf, L., Cohen-Or, D.: Optimizing photo composition. Computer Graphic Forum 29(2), 469–478 (2010)
14. Ahn, S., Agrawala, M., Hartmann, B., Barsky, B.A.: Image Cropping: Collection and Analysis of Crowdsourced Data Technical Report No. UCB/EECS-2012-94 (2012)
15. McManus, I.C., Zhou, F.A., l'Anson, S., Waterfield, L., Stöver, K., Cook, R.: The psychometrics of photographic cropping:The influence of colour, meaning, and expertise. Perception 40(3), 332–357 (2011)
16. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP) 7(1), Article 6 (2010)
17. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 12th International Conference on Computer Vision (2009)
18. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned Salient Region Detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 1597–1604 (2009)
19. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254 (1998)
20. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. Advances in neural information processing systems, vol. 19, pp. 545–552. MIT Press, Cambridge (2007)
21. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
22. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
23. Harris, C., Stephens, M.: A combined edge corner detector. In: 4th Alvey Vision Conference (1998)
24. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2), 353–367 (2011)