Automatic Image Cropping and Selection Using Saliency: An Application to Historical Manuscripts

Marcella Cornia(✉), Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
{marcella.cornia,stefano.pini,lorenzo.baraldi,
rita.cucchiara}@unimore.it

Abstract. Automatic image cropping techniques are particularly important to improve the visual quality of cropped images and can be applied to a wide range of applications such as photo-editing, image compression, and thumbnail selection. In this paper, we propose a saliency-based image cropping method which produces significant cropped images by only relying on the corresponding saliency maps. Experiments on standard image cropping datasets demonstrate the benefit of the proposed solution with respect to other cropping methods. Moreover, we present an image selection method that can be effectively applied to automatically select the most representative pages of historical manuscripts thus improving the navigation of historical digital libraries.

Keywords: Image cropping · Image selection · Saliency
Digital libraries

1 Introduction

Image cropping aims at extracting rectangular subregions of a given image with the aim of preserving most of its visual content and enhancing the visual quality of the cropped image [5,6,30]. A good image cropping algorithm can have several applications, from helping professional editors in the advertisement and publishing industry, to increasing the presentation quality in search engines and social networks, where it is often the case that variable sized images need to be previewed with thumbnails of given size. In the case of collections of images, the combination of frame selection and image cropping techniques can be exploited to generate high quality thumbnails representing the entire collection. The same line of thinking can be extended, of course, to the case of selecting appropriate thumbnail for a video.

Multimedia digital libraries, which contain collections of images and videos [2,4,13], are for sure a valuable application domain of image cropping and selection techniques. Motivated by these considerations, in this paper we devise a cropping technique based on saliency prediction. In fact, visual saliency prediction is

G. Serra and C. Tasso (Eds.): IRCDL 2018, CCIS 806, pp. 169–179, 2018.
https://doi.org/10.1007/978-3-319-73165-0_17

the task of predicting the most important regions of an image by identifying those regions which most likely attract human gazes at the first glance [10–12]. By relying on this information, we propose a simple and effective image cropping solution which returns cropped regions with the most important visual content of their corresponding original images. To validate the effectiveness of the proposed cropping technique, we assess its performance on standard image cropping datasets by comparing to state of the art methods.

Moreover, we propose an image selection method which exploits the ability of our cropping solution of finding the most important regions of images. In particular, to validate our solution in real-world scenarios, we apply it to the selection of the most representative pages of historical manuscripts. In this way, the selected pages can be used as an effective preview of each manuscript thus improving the navigation of historical digital libraries.

Overall, the paper is organized as follows: Sect. 2 presents the main related image cropping methods and briefly reviews the thumbnail selection literature, Sect. 3 introduces the proposed saliency-based cropping technique, while the corresponding experimental results are reported in Sect. 4. Finally, the automatic page selection of historical manuscripts is presented in Sect. 5.

2 Related Work

In this section, we start from reviewing the literature related to the automatic image cropping task. Also, we briefly describe some recent works addressing the thumbnail selection problem.

2.1 Image Cropping

Existing image cropping methods can be categorized into two main categories: *attention-based* and *aesthetics-based* methods. The first ones aim at finding the most visually salient regions in the original images, while the second ones accomplish the cropping task mainly by analyzing the attractiveness of the cropped image with the help of a quality classifier.

Attention-based approaches exploit visual saliency models or salient object detectors to identify the crop windows that more attract human attention [5,24, 26,27]. Some other hybrid methods employ a face detector to locate the regions of interest [32] or directly fit a saliency map from visually pleasurable photos taken by professional photographers [23]. Instead of using saliency, pixel importances can be also estimated using their objectness [9], or empirically defined energy functions [1,21].

On the other hand, aesthetics-based methods leverage on photo quality assessment studies [3,15,28] using certain objective aspects of images, such as low level image features and empirical photographic composition rules. In particular, Nishiyama *et al.* [22] built a quality classifier using low level image features such as color histogram and Fourier coefficient from which they selected the cropped region with the highest quality score. Chen *et al.* [8] presented a method to

learn the spatial correlation distributions of two arbitrary patches in an image for generating an omni-context prior which serve as rules to guide the composition of professional photos. Zhang *et al.* [31], instead, proposed a probabilistic model based on a region adjacency graph to transfer aesthetic features from the training photo onto the cropped ones.

More recently, Yan *et al.* [30] proposed several features that accounts the removal of distracting content and the enhancement of overall composition. The influence of these features on crop solutions was learned from a training set of image pairs, before and after cropping by expert photographers. Other works, instead, exploit a RankSVM [6], working with features coming from the AlexNet model [16], or an aesthetics-aware deep ranking network [7] to classify each candidate window. Finally, Li *et al.* [6] formulated the automatic image cropping problem as a sequential decision-making process, and proposed an Aesthetics Aware Reinforcement Learning (A2-RL) model to solve this problem.

2.2 Thumbnail Selection

The thumbnail selection problem has been widely addressed especially in the video domain, in which a frame that is visually representative of the video is selected and used as a representation of the video itself. In our case, instead, we want to find the most significant image from a collection of images (*i.e.* the pages of an historical manuscript), which somehow it can be considered as a related problem to the video thumbnail selection.

Most conventional methods for video thumbnail selection have focused on learning visual representativeness purely from visual content [14,20], while more recent researches have addressed this problem as the selection of query-dependent thumbnails to supply specific thumbnails for different queries.

Liu *et al.* [18] proposed a reinforcement algorithm to rank the frames in each video, while a relevance model was employed to calculate the similarity between the video frames and the query keywords. Wang *et al.* [29] introduced a multiple instance learning approach to localize the tags into video shots and to select query-dependent thumbnail according to the tags.

In [19], instead, a deep visual-semantic embedding was trained to retrieve query-dependent video thumbnails. In particular, this method employs a deeply-learned model to directly compute the similarity between the query and video thumbnails by mapping them into a common latent semantic space.

3 Automatic Image Cropping

We tackle the image cropping task as that of finding a rectangular region $\mathcal{R}$ inside the given image $\mathcal{I}$ with maximum saliency. Comparing to previous methods which maximized a function of the saliency inside $\mathcal{R}$, they all used other functions, such as the difference of saliency in $\mathcal{R}$ and outside $\mathcal{R}$, or the difference between the mean saliency value in $\mathcal{R}$ and the mean saliency value outside $\mathcal{R}$. We experimentally

validated that when using state of the art saliency predictors, our choice, although simple, provides better results than more fancy objective functions.

Formally, being $\mathbf{x}$ a pixel of the input image and $S(\mathbf{x})$ its saliency value, predicted by a saliency model, we aim at finding:

$$\max_{\mathcal{R}} \left(\int_{\mathbf{x} \in \mathcal{R}} S(\mathbf{x}) - \int_{\mathbf{x} \in \mathcal{I} \setminus \mathcal{R}} S(\mathbf{x}) \right) \tag{1}$$

This objective boils down to finding the minimum bounding box of all salient pixels, and taking all regions $\mathcal{R}$ which contains the minimum bounding box. Since taking regions larger than the minimum bounding box would amount to having non salient pixels in $\mathcal{R}$, we take $\mathcal{R}$ as the minimum bounding box of salient pixels.

Regarding the saliency map, we compute it for every image by using the saliency method proposed in [12] which currently is the state of the art method in the saliency prediction task. In particular, starting from a classical convolutional neural network, it iteratively refines saliency predictions by incorporating an attentive mechanism. Also, it is able to reproduce the center bias present in human eye fixations by exploiting a set of prior maps directly learned from data. Overall, the performance achieved by the selected saliency method allows us to rely on saliency maps that effectively reproduce the human attention on natural images.

4 Experimental Evaluation

In this section, we briefly describe datasets and metrics used to evaluate our solution and provide quantitative and qualitative comparisons with other image cropping methods.

4.1 Datasets

To validate the effectiveness of visual saliency in the automatic image cropping task, we perform experiments on two different publicly available datasets.

The Flickr-Cropping dataset [6] is composed of 1,743 images, each of them associated to ground-truth cropping parameters. Images are divided in training and test sets, respectively composed of 1,395 and 348 images. Our method is not trainable, but we perform experiments on test images only for a fair comparison with other methods.

The CUHK Image Cropping dataset [30] contains the cropping parameters for 950 images that were manually cropped by an experienced photographer. Images are provided with cropping annotations of three different photographers. In our experiments, we evaluate the performance of our saliency-based cropping method with respect to all three different annotations.

4.2 Metrics

Two different metrics are usually used to determine the accuracy of the automatic image cropping algorithms: the Intersection over Union (commonly abbreviated as IoU) and the Boundary Displacement Error (BDE).

The Intersection over Union is an evaluation metric used to evaluate the overlapping between two bounding boxes. Technically, it is defined as

$$\text{IoU} = \frac{1}{N}\sum_{i}^{N}\frac{GT_i \cap P_i}{GT_i \cup P_i} \tag{2}$$

where N is the number of samples, GT_i is the area of the ith ground-truth bounding box and P_i is the area of the ith predicted bounding box.

The Boundary Displacement Error measures the distance between the sides of the ground-truth bounding box and the predicted one. For convenience, the values are normalized with respect to the size of the image. Mathematically, the metric is defined as

$$\text{BDE} = \frac{1}{4}\frac{1}{N}\sum_{i}^{N}\left(\frac{|x_1^{GT_i} - x_1^{P_i}|}{w_i} + \frac{|y_1^{GT_i} - y_1^{P_i}|}{h_i} + \frac{|x_2^{GT_i} - x_2^{P_i}|}{w_i} + \frac{|y_2^{GT_i} - y_2^{P_i}|}{h_i}\right) \tag{3}$$

where N is the number of samples, (x_1, y_1) is the top left edge of the bounding box, (x_2, y_2) is the bottom right edge of the bounding box, w_i and h_i are respectively width and height of the image, GT_i is the ith ground-truth bounding box, and P_i is the ith predicted bounding box.

4.3 Results

We compare our solution with other automatic image cropping methods. For the Flickr-Cropping dataset, we perform comparisons with the most competitive saliency-based baseline presented in [6] (eDN), the RankSVM+DeCAF$_7$ model [6],

Table 1. Experimental results on the Flickr-Cropping [6] dataset. First, second and third best scores on each metric are respectively highlighted in red, green and blue colors.

Method	Avg IoU	Avg BDE
eDN [6]	0.4857	0.1372
RankSVM+DeCAF$_7$ [6]	0.6019	0.1060
VFN [7]	**0.6744**	**0.0872**
A2-RL [17]	**0.6564**	**0.0914**
Saliency density	0.6193	0.0997
VGG activations	0.6004	0.1088
Ours	**0.6589**	**0.0892**

the View Finding Network (VFN) proposed in [7] and the Aesthetics Aware Reinforcement Learning (A2-RL) model [17]. For the CUHK Image Cropping dataset, instead, the comparison methods are the change-based image cropping architecture presented in [30] (LearnChange) and the VFN and A2-RL models.

Moreover, for both datasets, we compare our results with two variations of our model which we call `Saliency Density` and `VGG Activations`. The first one aims at maximizing the difference of the averaged saliency between the selected bounding box and the outer region of the image. For simplicity, we set the size of search window to each scale among $[0.75, 0.80, \ldots, 0.95]$ of the original image and slide the search window over a 10×10 uniform grid. The `VGG Activations` is, instead, the proposed image cropping method where the saliency maps are replaced with the activations of the last convolutional layer of the VGG-16 network [25]. In particular, since the last convolutional layer has 512 filters, we select for each image the activation map having the maximum sum.

Table 1 shows the results on the Flickr-Cropping dataset. As it can be seen, our solution obtains the second best scores on both IoU and BDE metrics and achieves better results with respect to both our baselines. Table 2, instead, reports the results on the three different annotations of the CUHK Image

Table 2. Experimental results on three different annotations of the CUHK Image Cropping [30] dataset. First, second and third best scores on each metric are respectively highlighted in red, green and blue colors.

Annotation	Method	Avg IoU	Avg BDE
1	LearnChange [30]	0.7487	0.0667
	VFN [7]	**0.7847**	**0.0581**
	A2-RL [17]	**0.7934**	**0.0545**
	Saliency density	0.6345	0.0971
	VGG activations	0.7788	0.0574
	Ours	**0.8017**	**0.0500**
2	LearnChange [30]	0.7288	0.0720
	VFN [7]	**0.7763**	**0.0614**
	A2-RL [17]	**0.7911**	**0.0554**
	Saliency density	0.6053	0.1075
	VGG activations	0.7648	0.0624
	Ours	**0.7711**	**0.0594**
3	LearnChange [30]	0.7322	0.0719
	VFN [7]	0.7602	0.0653
	A2-RL [17]	**0.7826**	**0.0551**
	Saliency density	0.6153	0.1040
	VGG activations	**0.7612**	**0.0618**
	Ours	**0.7675**	**0.0599**

Cropping dataset. In this case, our method achieves the best results on the first annotation on both metrics, while, on the other two annotations, it obtains the second or the third best scores. Despite the proposed solution is much simpler than the other comparison methods, the results achieved by our method on both considered datasets are very close to the best ones, thus confirming the effectiveness of the proposed strategy. Finally, some qualitative results with the corresponding saliency maps are presented in Fig. 1.

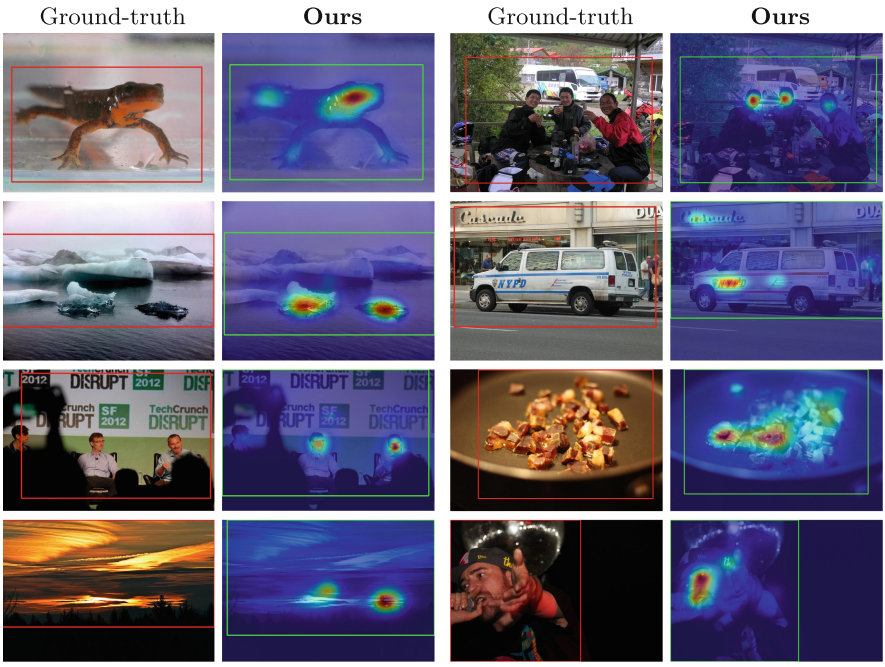

Fig. 1. Cropping results on sample images from the Flickr-Cropping dataset [6].

5 Automatic Page Selection of Historical Manuscripts

To validate our architecture in a real-world scenario, we apply it to find the best pages that represent historical manuscripts. This type of books usually have anonymous covers that does not represent its content, like plain colours or little artworks. Therefore, we develop a method to extract the most illustrative pages from every manuscript in order to use them as the preview of the book itself. Using this system, the navigation of historical digital libraries can be improved: users will be able to visually identify the content of a book watching its most representative images, without the need of opening it or read its summary.

In this case, the proposed image cropping method is not the output of the system, but it is used to find the most interesting pages of every manuscript.

In particular, the saliency map is calculated for every page of the book using the saliency model reported in [12]. After extracting all saliency maps,

Fig. 2. Example results of the page selection method on historical manuscripts. For each manuscript, the figure shows a list of some sample pages and the three pages selected by our method. As it can be seen, the selected pages contains representative visual contents and can be successfully used as a preview of the considered manuscript.

the method proposed in Sect. 3 is used to find the minimum crop that contains all the pixels with a saliency value higher than a threshold t (in our experiments $t = 128$). Then, a density score is calculated as the average value of saliency inside the bounding box divided by the average value of saliency outside the bounding box. In particular, it is formulated as

$$DS = \frac{\frac{1}{K}\sum_{i,j} s(i,j)}{\frac{1}{w \cdot h - K}\sum_{l,m} s(l,m)} \tag{4}$$

where K is the number of pixels inside the bounding box, (i, j) and (l, m) are respectively the coordinates of the pixels inside and outside the bounding box, while w and h are width and height of the image.

An high density score corresponds to an image where most of the saliency is restricted to a small area, therefore it contains a tiny region of high interest with respect to the rest of the image. On the contrary, a low density score corresponds to an image with a spread saliency map, therefore the image does not contain a valuable detail. Finally, the M images with the higher density score are selected as the most representative of the document.

Note that the method does not require training and it is applicable to any type of book, but it performs better with illustrated books. In our experiments, we decide to select entire images in place of image crops since we consider the full pages more suitable to be a summary of the whole manuscript, but it would be also possible to extract some particular details.

To validate our proposal, we apply the proposed automatic page selection method to a set of digitized historical manuscripts belonging to the Estense Library collection of Modena[1]. Some notable results are shown in Fig. 2. As it can be seen, the selected pages contain representative visual contents of the corresponding manuscript and they can be used as a significant preview of the manuscript itself.

6 Conclusions

In this work, we presented a saliency-based image cropping method which, by selecting the minimum bounding box that contains all salient pixels, achieves promising results on different image cropping datasets. Moreover, we applied our solution to the image selection problem. In particular, to validate the effectiveness in real-world scenarios, we introduced a page selection method which identifies the most representative pages of an historical manuscript. Qualitative results demonstrated that our idea improves the navigation of historical digital libraries by automatic generating significant book previews.

Acknowledgment. We gratefully acknowledge the Estense Gallery of Modena for the availability of the digitized historical manuscripts used in this work. We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

[1] http://bibliotecaestense.beniculturali.it.

References

1. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. ACM Trans. Graph. **26**(3), 10 (2007)
2. Balducci, F., Grana, C.: Affective classification of gaming activities coming from RPG gaming sessions. In: Tian, F., Gatzidis, C., El Rhalibi, A., Tang, W., Charles, F. (eds.) Edutainment 2017. LNCS, vol. 10345, pp. 93–100. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65849-0_11
3. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: ACM International Conference on Multimedia (2010)
4. Bolelli, F.: Indexing of historical document images: ad hoc dewarping technique for handwritten text. In: Grana, C., Baraldi, L. (eds.) IRCDL 2017. CCIS, vol. 733, pp. 45–55. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68130-6_4
5. Chen, J., Bai, G., Liang, S., Li, Z.: Automatic image cropping: a computational complexity study. In: IEEE International Conference on Computer Vision and Pattern Recognition (2016)
6. Chen, Y.L., Huang, T.W., Chang, K.H., Tsai, Y.C., Chen, H.T., Chen, B.Y.: Quantitative analysis of automatic image cropping algorithms: a dataset and comparative study. In: Winter Conference on Applications of Computer Vision (2017)
7. Chen, Y.L., Klopp, J., Sun, M., Chien, S.Y., Ma, K.L.: Learning to compose with professional photographs on the web. arXiv preprint arXiv:1702.00503 (2017)
8. Cheng, B., Ni, B., Yan, S., Tian, Q.: Learning to photograph. In: ACM International Conference on Multimedia (2010)
9. Ciocca, G., Cusano, C., Gasparini, F., Schettini, R.: Self-adaptive image cropping for small displays. IEEE Trans. Consum. Electron. **53**(4), 1622–1627 (2007)
10. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: International Conference on Pattern Recognition (2016)
11. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Multi-level net: a visual saliency prediction model. In: European Conference on Computer Vision Workshops (2016)
12. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an LSTM-based saliency attentive model. arXiv preprint arXiv:1611.09571 (2017)
13. Cucchiara, R., Grana, C., Prati, A.: Semantic transcoding for live video server. In: ACM International Conference on Multimedia (2002)
14. Kang, H.W., Hua, X.S.: To learn representativeness of video frames. In: ACM International Conference on Multimedia (2005)
15. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: IEEE International Conference on Computer Vision and Pattern Recognition (2006)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
17. Li, D., Wu, H., Zhang, J., Huang, K.: A2-RL: aesthetics aware reinforcement learning for automatic image cropping. arXiv preprint arXiv:1709.04595 (2017)
18. Liu, C., Huang, Q., Jiang, S.: Query sensitive dynamic web video thumbnail generation. In: IEEE International Conference on Image Processing (2011)
19. Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J.: Multi-task deep visual-semantic embedding for video thumbnail selection. In: IEEE International Conference on Computer Vision and Pattern Recognition (2015)

20. Luo, J., Papin, C., Costello, K.: Towards extracting semantically meaningful key frames from personal video clips: from humans to computers. IEEE Trans. Circ. Syst. Video Technol. **19**(2), 289–301 (2009)
21. Ma, M., Guo, J.K.: Automatic image cropping for mobile device with built-in camera. In: Consumer Communications and Networking Conference (2004)
22. Nishiyama, M., Okabe, T., Sato, Y., Sato, I.: Sensation-based photo cropping. In: ACM International Conference on Multimedia (2009)
23. Park, J., Lee, J.Y., Tai, Y.W., Kweon, I.S.: Modeling photo composition and its application to photo re-arrangement. In: IEEE International Conference on Image Processing (2012)
24. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semi-automatic photo cropping. In: SIGCHI Conference on Human Factors in Computing Systems (2006)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Stentiford, F.: Attention based auto image cropping. In: Workshop on Computational Attention and Applications, ICVS (2007)
27. Suh, B., Ling, H., Bederson, B.B., Jacobs, D.W.: Automatic thumbnail cropping and its effectiveness. In: ACM Symposium on User Interface Software and Technology (2003)
28. Tang, X., Luo, W., Wang, X.: Content-based photo quality assessment. IEEE Trans. Multimed. **15**(8), 1930–1943 (2013)
29. Wang, M., Hong, R., Li, G., Zha, Z.J., Yan, S., Chua, T.S.: Event driven web video summarization by tag localization and key-shot identification. IEEE Trans. Multimed. **14**(4), 975–985 (2012)
30. Yan, J., Lin, S., Bing Kang, S., Tang, X.: Learning the change for automatic image cropping. In: IEEE International Conference on Computer Vision and Pattern Recognition (2013)
31. Zhang, L., Song, M., Zhao, Q., Liu, X., Bu, J., Chen, C.: Probabilistic graphlet transfer for photo cropping. IEEE Trans. Image Process. **22**(2), 802–815 (2013)
32. Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W.: Auto cropping for digital photographs. In: ICME (2005)