Automatic product region extraction based on colour similarity and saliency detection models

Takuya Futagami and Noboru Hayasaka
Department of Engineering Informatics, Osaka Electro-Communication University, Neyagawa, Japan

ABSTRACT
In this paper, product region extraction, which can classify the pixels of the product images as product and background regions, is proposed. The proposed method is based on the handcrafted algorithm using both the colour similarity and the saliency detection. Our experiment, which employed 180 product images, clarified that the proposed method increased all the metric for the extraction accuracy compared with conventional methods based on the handcrafted algorithm. The F-measure, which is the comprehensive metric, was significantly increased by 2.20% or more. Our discussion also found that the proposed method also overcame the shortcoming of the conventional method, because the F-measure for the dataset, the accuracy of which was decreased by the conventional method, was significantly improved. In addition, the F-measure was increased by 0.92% or more for each product category. Further comparison and discussion are included in this paper to provide more focused findings.

1. Introduction
In recent years, consumer to consumer (C2C) online market has attracted international attention owing to widespread use of mobile devices such as smartphones and tablets [1]. Examples of C2C online marketplace include flea market (e.g. Mercari [2]) and auction (e.g. eBay [3]). At these marketplaces, sellers create product listing page so as to provide basic information such as name, category, manufacture and exterior appearance of their product to buyers. To reveal the current state of the product, sellers upload the product image, which is considerably important for the buyers to make the purchase decision [4].

Although the listing pages were conventionally created by text-based input from the sellers, several systems of online marketplaces have been reduced this burden by applying product identification to uploaded product image [5]. On the basis of the identified product information, the systems automatically create the listing pages. The product identification is mainly based on machine learning classifier such as Deep Neural Network (DNN) [6] and Support Vector Machine (SVM) [7]. To increase the identification performance, product region extraction, which can divide each pixel into product or background region, is expected to be employed so as to reduce noise affection from the background as discussed in [8]. In addition, the contour information provided by the product region extraction can be used as the identification feature [9,10].

Thus, the product region extraction is required to be improved so as to increase the identification accuracy.

The techniques of product region extraction are based on the handcrafted algorithm, which has been proposed on the basis of knowledge from researcher, and machine learning classifier. In recent years, machine learning classifier has been enthusiastically studied owing to the advent of semantic segmentation networks such as Fully Convolutional Network (FCN) [11] and SegNet [12]. However, the introduction cost of the machine learning classifier notably increased due to the requirement for training dataset with manual annotation at pixel-wise level [13]. On the other hand, the handcrafted algorithm can reduce the introduction cost due to no requirement for it.

In the literature, conventional product region extraction based on the handcrafted algorithm uses saliency detection, which quantitatively evaluates how much each pixel attracts the human gaze [14], or colour similarity. In [9], the saliency detection was applied to roughly determine the candidates of the product and background regions to initialize GrabCut [15], which can iteratively revise both the regions by creating and updating a graph on the basis of min–max flow algorithm [16] (hereafter called saliency-based method). The saliency model is used to predict product regions on the basis of an assumption that the product, which is the main subject of the image, tends to have a higher saliency value compared with
the background. As the saliency-based model, the saliency-based method employs global contrast-based model (hereafter called GC model) [17], which computes the saliency value for each superpixel, a group of pixels that have similar visual characteristics, created by Felzenszwalb method [18].

In [19], image boundary areas were assumed to be actual background region, and product and background candidates were roughly determined on the basis of the colour similarity from those areas (hereafter called colour-similarity-based method). The measurement of the colour similarity is conducted by applying region growing algorithm [20], which iteratively enlarges region by merging adjacent pixels from seed points.

An experiment in [19] demonstrated that the colour-similarity-based method increased extraction accuracy. However, our analysis, which is discussed in a later section, showed that the colour-similarity-based method can be improved by overcoming its shortcoming. Thus, this paper aims to increase the extraction accuracy of the handcrafted algorithm.

For this purpose, we propose a new method which is based on both the saliency detection and the colour similarity. The proposed method is organized to extend the background candidates properly compared with the colour-similarity-based method towards overcoming its shortcoming. To score the colour similarity, in addition to region growing algorithm, one class SVM [21], which is an unsupervised learning algorithm for detecting outlier on the basis of the Lagrange Multiplier formulation, is employed.

In the literature, the saliency detection model, which is effective for the product region extraction, had not been discussed. However, our previous work in [22] experimentally found that the most effective saliency model is manifold ranking based model (hereafter called MR model) [23]. This model applies Simple Linear Iterative Clustering (SLIC) method [24], which uses k-means clustering based on 5-D space of colour information and pixel location. On the basis of the work, the saliency model used in the proposed method was determined.

In Section 2 of this paper, we provide a description of related works of the handcrafted algorithm. In Section 3, we outline the proposed method. In Section 4, we report the experiment which aims to demonstrate the effectiveness of the proposed method thoroughly. We conclude in Section 5 with a summary.

2. Related work

This section outlines the conventional product region extraction based on the handcrafted algorithm. The procedures of the saliency-based method and colour-similarity-based method are presented in Sections 2.1 and 2.2, respectively.

2.1. Saliency-based method

In the first step, the GC model is applied to create the saliency map. Subsequently, the saliency map is thresholded at a predetermined level. The pixels which have higher saliency values are determined as the product candidates. The remained pixels are determined as the background candidates. Figure 1(b) depicts a saliency map from Figure 1(a). Figure 1(c) depicts the thresholded saliency map.

In the second step, GrabCut is applied. At each iteration, pixels which are not found at the border areas of both regions are detected. The detected pixels are allowed to be changed to the other region. Figure 1(d) depicts the extracted product from Figure 1(c).

2.2. Colour-similarity-based method

In the first step, region growing is applied to determine the product and background candidates. On the basis of the assumption that the background regions are found at the image boundary areas as discussed in [19], the colour-similarity-based method sets the seed points, which are depicted as grey pixels in Figure 2, to drive the region growing algorithm. The set of pixels in the enlarged region can be given as follows:

\[ R = \bigcup_{m \in M} R'_m \]  

where \( R'_m \) denotes the set of pixels included in the enlarged region provided by the single seed point \( m \). Figure 3(a) depicts the enlarged region from Figure 1(a). The pink and green areas represent the background and product candidates, respectively.

In the second step, GrabCut is applied. The background candidates are not allowed to be changed to the product region. Figure 3(b) depicts the extracted product from Figure 3(a).

Figure 1. Process of saliency-based method.
In [19], the colour-similarity-based method increased the accuracy compared with the saliency-based method. Hence, for further improvement, it is required to analyse the shortcoming of the colour-similarity-based method. Our analysis found that the colour-similarity-based method tended to incorrectly determine the actual background regions as the product region. This incorrect determination can be found in Figure 3(b). The other samples of this incorrect determination are depicted in Figure 4, which are obtained from Figure 5.

This incorrect determination is frequently found in the images which tend to contain image edges in the actual background regions, because image edges cause the region growing to misdetermine the actual background regions as the product candidate. To increase the accuracy, the actual background regions are required to be extensively determined as the background candidates.

2.3. Saliency detection model

The saliency-based method and proposed method use saliency detection model, the algorithm of which is developed by expert knowledge. The saliency-based method employs global contrast-based model (hereafter called GC model) [17], whereas the proposed method employs manifold ranking-based model (hereafter called MR model) [23]. Both models estimate the saliency value for each superpixel, which represents a group of pixels that have similar visual characteristics.

3. Proposed method

This section presents the proposed method. The algorithm is organized to extend the background candidates properly on the basis of measurement of the colour similarity from reliable candidates, which are outlined below, towards overcoming the shortcoming of the colour-similarity-based method discussed in Section 2.2. The flowchart of the proposed method is depicted in Figure 6, each step of which is presented in the subsequent subsections.

3.1. Determination of reliable product and background candidates

The first step determines the reliable product and background candidates. In the first procedure, the saliency map is created by applying the MR model to the input image. The MR model, details of which are outlined in [23], utilizes SLIC method to form the superpixels. The saliency value is computed for each superpixel on the basis of the manifold-based ranking according to relevance to the query superpixels which are found at top, bottom, left, and right image boundary. The
relevance is evaluated on the basis of the mean pixel values in Lab colour space for each superpixel. The created saliency map is used to determine the reliable background candidates on the basis of two types of thresholding. The first thresholding level is based on the predetermined value $t_s$. The second thresholding level is computed as follows:

$$t'_s = \bar{s} + \tau \times \sigma_s,$$  \hspace{1cm} \text{(2)}

$$\bar{s} = \frac{1}{Q} \sum_{q \in Q} s_q,$$  \hspace{1cm} \text{(3)}

$$\sigma_s = \sqrt{\frac{1}{Q} \sum_{q \in Q} s_q^2 - \bar{s}^2},$$  \hspace{1cm} \text{(4)}

where $Q$ denotes the set of pixels the saliency value of which is higher than $t_s$, $|Q|$ denotes the number of pixels included in the $Q$, $s_q$ denotes the saliency value at pixel $q$, and $\tau$ denotes the predetermined parameter to control the threshold level. The pixels which have higher saliency value than $t'_s$ are determined as the reliable product candidates.

In the second procedure, the region growing is applied the same setting as the colour-similarity-based method. The pixels in the enlarged region are set to the reliable background candidates. Figure 7(a) presents the superpixels provided by applying the MR model to Figure 1(a). The yellow lines represent the borders of superpixels. In Figure 7(b), the superpixels which are located not at black areas are employed as query for the manifold-based ranking. For each superpixel, the MR model computes the saliency value as depicted in Figure 7(c). The green and pink areas of Figure 7(d) represent the reliable product and background candidates, respectively.

### 3.2. Determination of product and background candidates

The second step determines the product and background candidates (but not reliable ones) from the remained pixels, which were not determined as the reliable candidates in the previous subsection.

To overcome the shortcoming of the colour-similarity-based method, it is expected to be effective that the background candidates which are not found near the image boundary areas are prepared. Thus, the one class SVM, in which the pixels in the reliable background candidates are employed as the training pixels, is applied. The scores are used as the measurement of the colour similarity to the reliable background candidates. The pixels the score of which is higher than $t_0$ as follows, are determined as the background candidates:

$$t_0 = \frac{1}{|Q'|} \sum_{q \in Q'} o_q,$$  \hspace{1cm} \text{(5)}

where $Q'$ denotes the set of pixels which were not determined as reliable candidates in the previous step, and $o_q$ denotes the score computed by the one class SVM for pixel $q$. Otherwise, the pixels are determined as the product candidates. The blue and grey areas of Figure 7(d) represent the product and background candidates, respectively.
3.3. Apply GrabCut

The third step applies GrabCut initialized by the candidates determined in the previous steps. The reliable candidates are not allowed to be changed to the other region, whereas the other candidates are allowed. Figure 7(e) depicts the extracted product from Figure 7(d).

The comparison between Figure 3(b) and Figure 7(e) suggests that the proposed method overcame the shortcoming of the colour-similarity-based method. In addition, this suggestion is supported by the fact that Figure 8, which is also obtained from Figure 5, tended to overcome the shortcoming compared with Figure 4. Hence, because the proposed method is expected to improve the accuracy, the experiment is conducted in the next section.

4. Experiment

In this section, we evaluate the effectiveness of the proposed method. Sections 4.1 and 4.2 present the experimental conditions and results, respectively. Section 4.3 discusses the effectiveness of the proposed method via analysis from different viewpoints.

4.1. Experimental condition

This subsection presents the experimental conditions. Sections 4.1.1 and 4.1.2 outline the evaluation dataset and metric, respectively. Section 4.3 discusses the comparative methods for the proposed method.

4.1.1. Evaluation dataset

For evaluation, 180 product images were captured by us, and their ground truth image, the product and background of which are correctly specified, were manually prepared. The samples of this evaluation dataset are listed in Figure 9. The evaluation dataset can be divided into nine product categories, which are set based on [19] (fashion, home, video game/book/music, hobby, beauty, electronics, sport, handmade, and other). Each category has 20 images.

4.1.2. Evaluation metric

This experiment evaluated the proposed method in terms of product extraction accuracy based on precision, recall, and F-measure, which are defined as:

\[
\text{Precision} = \frac{\text{True - positive}}{\text{True - positive} + \text{False - positive}}
\]

\[
\text{Recall} = \frac{\text{True - positive}}{\text{True - positive} + \text{False - negative}}
\]

\[
\text{F - measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

where the true-positive and false-positive denote the number of correct and incorrect determinations as the product. In addition, the false-negative denote the number of incorrect determinations as the background. The precision and recall are in an unavoidable trade-off in general, the F-measure, which is their harmonic mean, is employed as the comprehensive metric.

4.1.3. Comparative method

The proposed method was compared with the saliency-based method, the colour-similarity-based method, low-cost SegNet, and an interactive GrabCut. The saliency-based method, the colour-similarity-based method, and the proposed methods are categorized as the handcrafted algorithm presented in Sections 2.1, 2.2, and 3, respectively. The low-cost SegNet is categorized as the machine learning classifier.

The parameters of the saliency-based method and the colour-similarity-based method were determined experimentally on the basis of preliminary experiment using 300 product images, which are not included in the evaluation dataset. This preliminary dataset was used to create the colour-similarity-based method in [19], where the details of this dataset are outlined. SegNet,
which requires a large-scale annotated training dataset to enhance the accuracy, was trained with this dataset, which may not include a sufficient amount of various product images. These conditions were followed from [19].

The employment of SegNet stems from the fact that it has been the only DNN applied to the product extraction in the literature. The encoder was pretrained by large-scale dataset ImageNet, which has million images, to improve the generalization ability. Other conditions for training were followed from [12].

Interactive GrabCut, which is based on [15], applied GrabCut initialized by the bounding box, which is specified manually. The outer and inner areas are determined as the background candidates and product candidates, respectively. The product candidates are allowed to be changed to the background regions, whereas the background candidates are not allowed, when GrabCut is applied. Figure 10(b) is the extracted product based on the green bounding box in Figure 10(a). Although this method does not automatically extract the product regions, it can be one of a measurement to grasp a room for improvement of the proposed method. The bounding boxes were prepared on the basis of the ground truth images.

4.2. Experimental results

Table 1 summarizes the extraction accuracy with each method. We analysed this table on the basis of Wilcoxon signed-rank test, which is the non-parametric paired comparative test, to detect significance differences at a significance level 5%.

Compared with the saliency-based method and colour-similarity-based method, the proposed method increased all the metrics. The precision was non-significantly increased by 0.80% or more, whereas the recall and F-measure were increased by 3.55% or more and 2.20% or more, respectively. These results suggest that the shortcoming of the colour-similarity-based method, which decreased the recall, was overcome by the proposed method.

Compared with the low-cost SegNet, the saliency-based method significantly decreased the F-measure by 3.38%, and the colour-similarity-based method nonsignificantly increased the F-measure by 0.38%. However, the proposed method was effective in that it could significantly increase the F-measure by 2.58%. This result indicates that the proposed method was effective compared with the low-cost SegNet trained by small-scale dataset to estimate the parameters for the handcrafted algorithms. However, we should note that the accuracy of SegNet is expected to be improved by preparing a various training dataset.

Compared with the interactive GrabCut, the other methods decreased all the metrics. The saliency-based method, colour-similarity-based method, and the low-cost SegNet significantly decreased the F-measure by 3.71% or more. However, there was no significant difference between the proposed method and the interactive GrabCut in the F-measure. These results suggest that the proposed method can be further improved by modifying the algorithm to create the candidates. However, the proposed method was effective in that it could automate the product region extraction without significantly decreasing the accuracy.

Figure 11 compares the extracted product regions with each method for three samples, the F-measure of which is evaluated in Table 2. At the first sample, the figures show that the colour-similarity-based method, the proposed method, and the interactive GrabCut could extract the product almost the same as the ground truth images, as demonstrated by their considerable F-measure of 96.10% or more. At the second sample, the proposed method and the interactive GrabCut increased the F-measure by 5.28% or more compared with the colour-similarity-based method. This result is provided by the fact that these methods tended to correctly determine the actual background regions as the background as depicted in the figures. At the third sample, the proposed method increased the F-measure by 3.94% or more compared with the colour-similarity-based method and the interactive GrabCut, because it tended to avoid determining the actual product regions as the background. This effect stems from the fact that the proposed method determines the reliable product candidates, which are not allowed to be changed to the background, on the basis of saliency values. On the other hand, the low-cost SegNet tended to incorrectly determine the actual background regions as the product at all the samples. SegNet may require the various training dataset to enhance its generalization ability.
4.3. Discussion

This subsection discusses the effectiveness of the proposed method to provide more focused findings. The Wilcoxon signed-rank test is also used to detect the significance differences.

We divided the evaluation dataset into two subsets: low precision dataset and low recall dataset, which are grouped on the basis of the accuracy of the colour-similarity-based method. The low-precision dataset contains 66 images the precision of which was lower than recall, whereas the low-recall dataset contains the other 114 images. We should note that the low-recall dataset tends to include images the recall of which was decreased by the shortcoming of the colour-similarity-based method.

Table 3 summarizes the extraction accuracy for the low-precision dataset and the low-recall dataset. For both the datasets, the proposed method significantly increased the F-measure compared with the saliency-based method, the colour-similarity-based method, and the low-cost SegNet. This result supports the fact that the shortcoming of the colour-similarity-based method was overcome as expected. In addition, the proposed method was effective for the other images.

Table 4 summarizes the F-measure for each product category. Compared with the saliency-based method and the low-cost SegNet, the proposed method significantly increased the F-measure by 0.97% or more for all the categories. Compared with the colour-similarity-based method, the F-measure was non-significantly increased by 0.92% for video/book/music category.

However, because the interactive GrabCut also did not increased the F-measure for this category, the colour-similarity-based method and the proposed method might achieve considerable accuracy for this category. On the other hand, the proposed method significantly increased F-measure by 1.40% or more for the other categories. Compared with the interactive GrabCut, the F-measure tended to be decreased. However, there were no significant differences for all the categories. Although these results support the effectiveness of the proposed method, we should note that each category includes only 20 product images.
Although the proposed method could enhance the accuracy compared with the conventional methods, the further improvement is required for images, in which there are no considerable differences in the colour between product and background, as depicted in Figure 12. For both the images, the white segments of the actual product, which are misclassified as the background candidates due to the colour similarity, tend to be misdetermined as the background. This misdetermination, which decreases the recall, may be solved by employing an information of the contour structure of the product in the future.

5. Conclusions

In this paper, the product region extraction based on the handcrafted algorithm was proposed. The proposed method employs the techniques for both the colour similarity and the saliency detection, which are also used in the conventional methods. This method aims to extend the background candidates properly compared with the conventional method so as to overcome its shortcoming.

The experiment, which is based on 180 product images, demonstrated that the proposed method significantly increased the F-measure, comprehensive metric for the extraction accuracy, by 2.20% or more compared with the conventional product region extraction based on the handcrafted algorithm. The F-measure was also increased by 3.38% compared with the low-cost SegNet, which was trained by small-scale dataset to estimate the parameters for the handcrafted algorithms. However, the F-measure was decreased by 1.51% compared with the interactive GrabCut, which is initialized by manual input. Although this comparison found the effectiveness of the proposed method for each comparative method, further room for the improvement was also detected.

Our discussion, which aims to provide more focused findings, clarified that the proposed method overcame the shortcoming of the conventional method. This statement stems from the fact that the F-measure was significantly increased for the low-recall dataset, the accuracy for which was decreased by that shortcoming. For each product category, the effectiveness of the proposed method could be found, since the F-measure was increased by 0.92% or more compared with the conventional methods and low-cost SegNet.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Notes on contributors

Takuya Futagami received his BS, MS, and PhD degrees from Osaka University in 2013, 2015, and 2021, respectively. He has been a visiting researcher at Osaka Electro-Communication University since 2018. At present, he is a lecturer at Aichi Gakuin University, Nishin, Japan. His current research interests include image processing and image recognition.

Noboru Hayasaka received his BS, MS, and PhD degrees from Hokkaido University, Sapporo, Japan, in 2002, 2004, and 2007, respectively. At present, he is an associate professor at Osaka Electro-Communication University, Neyagawa, Japan. His current research interests include speech processing and recognition.

References

[1] Kodate S, Chiba R, Kimura S, et al. Detecting problematic transactions in a consumer-to-consumer e-commerce network. Appl Netw Sci. 2020;5(1):1–18.
[2] Haghirian P. A Japanese start-up conquering the world – Mercari. In: Haghirian P, editor. Japanese management: market entry, crisis and corporate growth. Singapore: World Scientific; 2021. p. 17–29.
[3] Li D, Li J, Lin Z. Online consumer-to-consumer market in China – a comparative study of Taobao and eBay. Electron Commer Res Appl. 2008;7(1):55–67.
[4] Ow TT, Spaid BI, Wood CA, et al. Trust and experience in online auctions. J Org Comput Electron Commer. 2018;28(4):294–314.
[5] Fuchs G, Roitman H, Mandelbrod M. Automatic form filling with form-BERT. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Montreal, Canada. 2021. p. 1850–1854.
[6] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: a brief review. Comput Intell Neurosci. 2018;2018:ArticleID 7068349.
[7] Chandra MA, Bedi SS. Survey on SVM and their application in image classification. Int J Inf Technol. 2021;13:1–11.
[8] Fang W, Ding Y, Zhang F, et al. DOG: a new background removal for object recognition from images. Neurocomputing. 2019;361:85–91.
[9] He J, Feng J, Liu X, et al. Mobile product search with bag of hash bits and boundary reranking. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA. 2012. p. 3005–3012.
[10] Shen X, Lin Z, Brandt J, et al. Mobile product image search by automatic query object extraction. Proceedings of the European Conference on Computer Vision, Florence, Italy. 2012. p. 114–127.
[11] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–651.
[12] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–2495.

[13] Zhang M, Zhou Y, Zhao J, et al. A survey of semi-and weakly supervised semantic segmentation of images. Artif Intell Rev. 2020;53(6):4259–4288.

[14] Ullah I, Jian M, Hussain S, et al. A brief survey of visual saliency detection. Multimed Tools Appl. 2020;79(45):34605–34645.

[15] Rother C, Kolmogorov V, Blake A. Grabcut: interactive foreground extraction using iterated graph cuts. ACM Trans Graph. 2004;23(3):309–314.

[16] Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans Pattern Anal Mach Intell. 2004;26(9):1124–1137.

[17] Cheng MM, Mitra NJ, Huang X, et al. Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell. 2015;37(3):569–582.

[18] Felzenszwalb PF, Daniel PH. Efficient graph-based image segmentation. Int J Comput Vision. 2004;59(2):167–181.

[19] Futagami T, Hayasaka N. Automatic product region extraction based on analysis of images uploaded to C2C online market. J Org Comput Electron Commer. 2020;30(4):323–334.

[20] Adams R, Bischof L. Seeded region growing. IEEE Trans Pattern Anal Mach Intell. 1994;16(6):641–647.

[21] Alam S, Sonbhadra SK, Agarwal S, et al. One-class support vector classifiers: A survey. Knowl Based Syst. 2020;196:105754.

[22] Futagami T, Hayasaka N. Experimental evaluation of saliency detection model for product region prediction. Proceedings of the 2021 60th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Tokyo Japan. 2021. p. 3166–3173.

[23] Yang C, Zhang L, Lu H, et al. Saliency detection via graph-based manifold ranking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, USA. 2013. p. 3166–3173.

[24] Achanta R, Shaji A, Smith K, et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell. 2012;34(11):2274–2282.