# Text Localization in Born-Digital Images of Advertisements

Dirk Siegmund(✉), Aidmar Wainakh, Tina Ebert, Andreas Braun, and Arjan Kuijper

Fraunhofer Institute for Computer Graphics Research (IGD), Fraunhoferstraße 5, 64283 Darmstadt, Germany
{dirk.siegmund,aidmar.weinakh,tina.ebert,andreas.braun, arjan.kuijper}@igd.fraunhofer.de

**Abstract.** Localizing text in images is an important step in a number of applications and fundamental for optical character recognition. While born-digital text localization might look similar to other complex tasks in this field, it has certain distinct characteristics. Our novel approach combines individual strengths of the commonly used methods: stroke width transform and extremal regions and combines them with a method based on edge-based morphologically growing. We present a parameter-free method with high flexibility to varying text sizes and colorful image elements. We evaluate our method on a novel image database of different retail prospects, containing textual product information. Our results show a higher f-score than competitive methods on that particular task.

## 1 Introduction

Optical character recognition (OCR) is a widely used application describing the task of converting images of typed, handwritten or printed text into machine readable text. One challenge in OCR is the localization of text within an image. While born-digital (BD) text localization might look similar to other complex tasks of this field, they present certain distinct characteristics. They originate from materials in digital form and carry salient semantic information such as advertisements or security-related information. Therefore, extracting text information from BD images enhances the semantic relevance of its content for indexing and retrieval. One application example is retail advertisement, where brochures and prospects are the most common marketing media elements. To make their offline advertisements accessible on new media devices like tablets and smart-phones, internet platforms digitize and index them [1,8]. An important component in this challenging task is the localization of text on these images (see Fig. 1). They are available free of charge to readers as a printed edition or as e-paper on the internet. Mainly they contain product names, with its textual description, price and images. Usually, OCR systems extract these kind of text by following the steps: text localization, text segmentation and text recognition. Text localization is critical to the overall system performance and is influenced by varying image background, text color, the used font and layout (see Sect. 3).


M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 627–634, 2018.
https://doi.org/10.1007/978-3-319-75193-1_75

When BD images are getting used, the performance of known OCR methods drop drastically. Main difficulty in theses methods are the definition of color clusters and parameters for local thresholding. We found additional challenges and features that reduced the appliance of state of the art methods: extremal regions (ER) and stroke width transform (SWT) to the recognition of BD advertisement images. To overcome these problems, we (1) provide a new technical viewpoint for the localization of text in born-digital advertisement images. Our method combines the strength of SWT and ER with a novel method based on morphology growing and edge-detection. (2) We present an entropy based color clustering method that is flexible to colorful text elements. (3) A novel database is presented, containing advertisement prospects from different retailers.

In Sect. 2 we will give a brief overview about related methods. An overview of our database is given in Sect. 3, where we also focus on the special properties of our database and the differences to other data-sets of BD images. Our method is introduced in Sect. 4, where we explain, how we addressed the identified challenges. In Sect. 5 we show our results in comparison with other methods.

## 2 Related Work

Research on BD images is focused mostly on applications like image-spam filtering, trying to classify images in emails as legit or spam [5]. In these tasks defining the extent of text is more important, then recognizing its content [6]. We found following methods developed for these domains not necessary transferable to our task. Nevertheless, known general techniques on text localizing are the detection of ER across several image projections (channels), as proposed by Neumann and Matas [10]. In their approach, a two phase classifications is presented using the features: area, bounding box, perimeter and Euler number for classification with AdaBoost. In the second phase a SVM classifier is used for classification of the hole area ratio, convex hull ratio, and the number of outer boundary inflexion points as features. Moreover, a clustering algorithm groups ER's into text lines, where finally an OCR classifier recognizes the text. Epshtein [4] and proposed using SWT to detect text in natural scenes images. A method proposed by Cho et al. [3] exploited the similarity between characters in order to detect text. Their approach is based on the fact that the cohesive characters compose a word or sentence are sharing similar properties such as spatial location, size, color, and stroke width. It consists of finding ER's, applying double threshold classifier and hysteresis tracking. This work achieved a recall harmonic mean of 82.17%, exceeding the winner in ICDAR 2013 RRC (Challenge 2). Yu et al. presented an approach [13] that localizes text, based on edge analysis. The authors used over-segmentation of edges and recombined them using a neighbor map, where they applied an edge filter, using stroke width, angle between corresponding points, shape feature, and the variance of morphological operation. Moreover, a text line filter was applied, which considers that characters in the same line having similar color, stroke width, height, and width.

## 3 Database

BD images are digitally created images that may show any content as textual information. As many of them are used online, low resolution and compression artifacts (see Fig. 1c) may occur. Especially in advertisements, there is a variety of text sizes and fonts (see Fig. 1e and f), which sometimes have complex and unusual designs. In comparison to real-scene images, the amount of text can be quite large and often contains text that is of no particular interest (like brand names on the pictured products as in Fig. 1d). Sometimes, the background is a complex image with high contrast, which makes the text detection more difficult (see Fig. 1a). Also, the text is spread over the whole page in small groups and is not placed in one well defined block, as it is in pure text documents (see Fig. 1b). Another challenge is the variety of colors, although it is often confined to a maximum number of three because of design reasons. This is helpful in determining the number of required color clusters. Other beneficial characteristics are the usually high text to background contrast and the good readability, which by contrast is challenging in real-scene images. Also, there are no illumination problems and no perspective transformations.


**Fig. 1.** (a) High contrast background (b) different alignment (c) compression artifacts (d) undesired text elements (e) variety of text sizes (f) special fonts. (Color figure online)

### 3.1 Advertisement Database

We created a novel advertisement database, containing a number of 78 different advertisement prospects. This included advertisements from local retailers, like beverage shops (see Fig. 1 left) and supermarkets with challenging content. In all cases, the total image size is 1200 × 1860 pixel (px) and a combination of

product pictures, product information and prizes is shown. The layout differs for every page, and there are both, single color backgrounds and complex figure backgrounds. A large variety in text sizes, fonts and colors is also included on each page. The ground truth was created manually by creating bounding boxes, covering text elements line-wise and drawing them on a binary mask.

## 4 Our Approach

We present a novel approach for the localization of text in images of advertisement prospects. We combined useful aspects of the three known text localization methods: Stroke Width Transform [4] and Extremal Regions [9] and Edge Based Morphological Growing [7] into a novel method in order to localize the different text elements on the images of our data-set.

### 4.1 Edge-Based Morphological Growing

This approach uses morphological operators in order to merge nearby detected edges to form connected components [7].

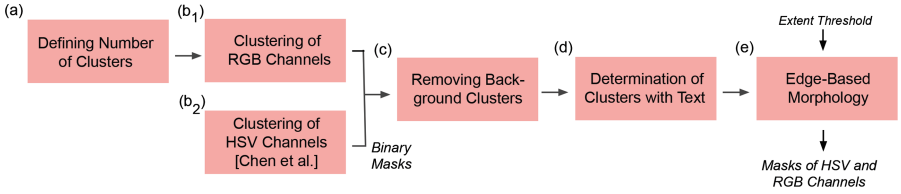


**Fig. 2.** Process of edge-based morphological growing.

Advertisement images are most often very colorful but still provide good readability and contrast of the relevant information. But in our tests, analyzing advertisements of many companies, we noticed that relevant product information like: product name, price and its description show not more than two or three colors. Thus, we cluster the image colors in two different ways by using k-mean to reduce the complexity of the image (Fig. 2$b_1$ and $b_2$). As the use of different color models represents different color-perception, we use RGB and HSV color model and combine their advantages in later processing.

1. Entropy based Color Clustering (RGB): Images contain different amount of colors, which is reflected by the image entropy. In Formula 1 we propose using the 'Shannon entropy' to define the amount of $k$ color-clusters. First, we convert the image to HSV color space and calculate the entropy $E$ of the H channel. Second, we assign $E$ divided by 2 to the formula on the right, to find

$k$ number of clusters where $\alpha = 0.2$ and $\beta = 2.4$. The variables $\alpha$ and $\beta$ are defined by testing iteratively on our data-set.

$$E(I) = -\sum_{C=1}^{C=3}\sum_{i=1}^{i=2^8} p(_i)\log_2(p(_i)), k = e^{\left(\frac{E(I)-\alpha}{\beta}\right)} \quad (1)$$

2. Centroid based Color Clustering (HSV): We use the approach proposed by Chen et al. [2] operating across different dimension of a HSV image. The initialization of centroids and thereby the number of clusters is calculated using histogram quantization.

After having clusters from these two methods we exclude their main background cluster (see Fig. 2c) by choosing the one with most non zero pixel. Thereafter, we calculate the amount of text in each cluster using the software 'Google Tesseract' [12]. We choose the three clusters with the highest amount of text (Fig. 2d) for further processing. In the last processing step (see Fig. 2e) we detect edges using the 'Sobel Edge Detector' and apply morphological closing horizontally. Since characters are rich of edges, the closing operation will lead to merging them. The size of the closing operation element was chosen by testing over the full data-set. Finally, we detect connected components as candidates for text lines using structural component analysis. Since the output of this approach contains a lot of false positive, we filter components based on their geometric properties extent and size. Second, we filter smaller isolated boxes, since most likely text is gathered in lines and blocks. In the last step binary masks are generated from both, the RGB and HSV clustered masks containing the remaining components.

### 4.2 Stroke Width Transform

Stroke width transform has shown its advantages in different applications of this field [11]. We use the approach of Epshtein [4] because it represents a suitable image operator representing the stroke line of text. To define stroke width for each image pixel and group pixel that have similar stroke width. Each group of pixels is considered as a candidate letter because usually letters have same stroke width. Non-letter groups get filtered out to aggregate text elements. SWT shows limitation in detecting especially the smaller characters. We extended this method by a measurement of the intensity of characters in a specific position in the image. Based on that, we decide whether there is a missed character at that position. Most likely the missed character is surrounded by several detected characters within the same line or the upper/lower lines.

### 4.3 Extremal Regions

We employed the ER text localization method proposed by Neumann and Matas (NM) [9], as they showed good results on similar tasks. A corresponding program flow chart is given in Fig. 3, where in (a), a combination of red (R), green (G),

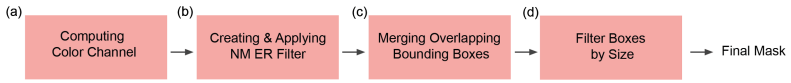


**Fig. 3.** Sequence of ER detection.

blue (B), and an intensity gradient ($\nabla$) is used to calculate channels, that are to be processed individually.

In (b) 1st and 2nd stage classifiers are generated. In the first stage, extremal regions are detected by using incrementally computable descriptors, which are then classified as text or non-text in the second stage. These two ER filters are both applied to the previous calculated NM channels, and return ER objects and their bounding boxes. Afterward, in the overlapping boxes are merged to avoid redundancy (Fig. 3c). Especially when trying to localize text boxes of similar size, such as prices, the previously found boxes need to be filtered (d). For this, the most occurring box height is determined by using a histogram of heights. Then, the average area of boxes with a height around this value is calculated, and used to filter out over- and undersized boxes. In the last step a binary mask is created from the remaining set of bounding boxes.

### 4.4 Combination of Masks

We integrate these methods, using their individual strength on different text sizes. We distinguish between: small (<60px height) and big (>60px height) texts elements. We use ER to detect the big text, because it shows the highest amount of false-positive while we used morphology to detect small text. In addition, SWT is used to detect any text size since it is not text size sensitive.

**First combination:** In our experiments, we noticed that the masks resulting from 'Edge-Based Morphological Growing' RGB and HSV clustering rarely contain the same non-text boxes. Therefore, we intersect elements on both masks and verify them by using the SWT mask. Boxes that were excluded intersecting RGB and HSV mask, get included again, in case they are found in SWT mask. Then we add the ER mask since it is responsible of detecting big texts.

**Second combination:** In this combination we used SWT as a base mask and added to it the ER mask, containing the big text elements. As SWT has limitation in detecting small texts, we add a intersection mask of the RGB and HSV resulting from the 'Edge-Based Morphological Growing' method.

## 5 Results

The evaluation was performed in terms of text element-level localization on our data-set. To quantify our results we used three parameters: recall rate, precision, and f-score. ER led to very high false positive rate as shown in Table 1. The number of the detected elements was 5795, which is more than double of the

**Table 1.** Results of proposed combination of methods.

| Approach | #Elements | Recall% | Precision% | f-score% |
|---|---|---|---|---|
| RGB | 3736 | 69.95 | 50.98 | 57.91 |
| HSV | 3992 | 52.72 | 52.40 | 50.29 |
| SWT | 1788 | 73.84 | 70.95 | 71.44 |
| ER | 5795 | 71.04 | 50.35 | 57.95 |
| Combination 1 | 2732 | 75.21 | 65.53 | 68.88 |
| Combination 2 | 2005 | 88.54 | 67.63 | 75.91 |

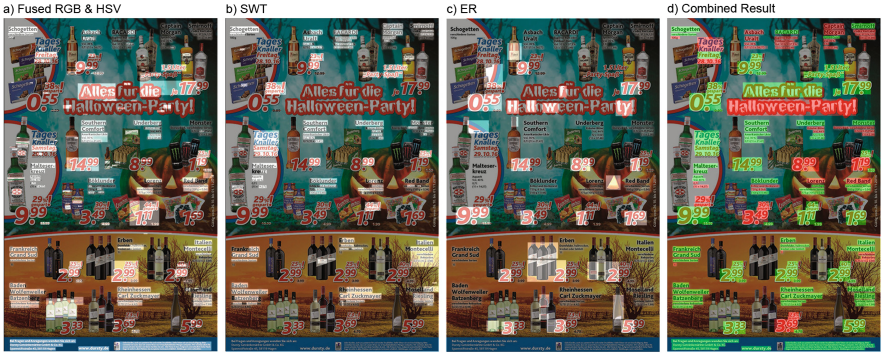


**Fig. 4.** Example, showing individual results (a–c) and final results (d).

ground truth elements of 2077. Our evaluation showed that most of these falsely detected elements are of small size, while the approach achieved better detection and precision rate for big elements. Edge-Based Morphological Growing showed its strengths especially in detecting areas of small font size. We noticed low recall rate and low precision, due to the limited filtering criteria that is available in this approach when using both RGB and HSV clusters. The main criteria to filter out the non-text blobs was the extent, which is the ratio between the size, white pixel and their bounding box. SWT reached a high recall rate in case of big text while its performance decreases when the text becomes smaller. We presented two combinations of the presented methods, as shown in Table 1. In combination 1, the morphology approach is mainly used to detect small texts but has a high false positive rate. Therefore more elements were detected in combination 1 compared to combination 2. On the other hand, we noticed better precision in combination 2 as it depends on SWT, which strictly verifies the intersection of the masks gathered by the edge-based morphology growing method (Fig. 4).

## 6 Conclusion

We presented a novel localization method for born-digital images. We showed the characteristics of this specific use-case and explored methods that are used in

similar tasks. Our presented method adapts several strengths of these methods and extends them to be more robust against challenges of advertisement images. We introduced a novel image database, containing advertisement prospects on which we evaluated our method. As a result we showed that our method overcomes limitation of competitive methods. Our pipeline is able to detect text, without any parameter, regardless of its scale, color and font.

**Acknowledgment.** This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within CRISP.

## References

1. Bonial International GmbH: Kaufda (2017). http://www.kaufda.de/
2. Chen, T.W., Chen, Y.L., Chien, S.Y.: Fast image segmentation based on K-Means clustering with histograms in HSV color space. In: 2008 IEEE 10th Workshop on Multimedia Signal Processing, pp. 322–325 (2008)
3. Cho, H., Sung, M., Jun, B.: Canny text detector: fast and robust scene text localization algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3566–3573 (2016)
4. Epshtein, B.: Detecting text in natural scenes with stroke width transform, pp. 2963–2970 (2010)
5. Gonzalez, A., Bergasa, L.M., Yebes, J.J., Bronte, S.: Text location in complex images. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 617–620. IEEE (2012)
6. Hanif, S.M., Prevost, L.: Text detection and localization in complex scene images using constrained adaboost algorithm. In: 2009 10th International Conference on Document Analysis and Recognition, ICDAR 2009, pp. 1–5. IEEE (2009)
7. Khan, N., Puri, S.: A study on text detection techniques of printed documents. In: Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016, pp. 2478–2482 (2016)
8. marktguru Deutschland GmbH: Marktguru (2017). http://info.marktguru.de/
9. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3538–3545. IEEE (2012)
10. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1872–1885 (2016)
11. Siegmund, D., Ebert, T., Damer, N.: Combining low-level features of offline questionnaires for handwriting identification. In: Campilho, A., Karray, F. (eds.) ICIAR 2016. LNCS, vol. 9730, pp. 46–54. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41501-7_6
12. Smith, R.: An overview of the Tesseract OCR engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, ICDAR 2007, Washington, DC, USA, vol. 02, pp. 629–633. IEEE Computer Society (2007)
13. Yu, C., Song, Y., Zhang, Y.: Scene text localization using edge analysis and feature pool. Neurocomputing **175**, 652–661 (2016)