A Text Recognition and Retrieval System for e-Business Image Management

Jiang Zhou ✉, Kevin McGuinness, and Noel E. O'Connor

Dublin City University, Dublin, Ireland,
{jiang.zhou, kevin.mcguinness, noel.oconnor}@dcu.ie

Abstract. The on-going growth of e-business has resulted in companies having to manage an ever increasing number of product, packaging and promotional images. Systems for indexing and retrieving such images are required in order to ensure image libraries can be managed and fully exploited as valuable business resources. In this paper, we explore the power of text recognition for e-business image management and propose an innovative system based on photo OCR. Photo OCR has been actively studied for scene text recognition but has not been exploited for e-business digital image management. Besides the well known difficulties in scene text recognition such as various size, location, orientation in text and cluttered background, e-business images typically feature text with extremely diverse fonts, and the characters are often artistically modified in shape, colour and arrangement. To address these challenges, our system takes advantage of the combinatorial power of deep neural networks and MSER processing. The cosine distance and n-gram vectors are used during retrieval for matching detected text to queries to provide tolerance to the inevitable transcription errors in text recognition. To evaluate our proposed system, we prepared a novel dataset designed specifically to reflect the challenges associated with text in e-business images. We compared our system with two other approaches for scene text recognition, and the results show our system outperforms other state-of-the-art on the new challenging dataset. Our system demonstrates that recognizing text embedded in images can be hugely beneficial for digital asset management.

Keywords: image management, image retrieval, OCR

1 Introduction

As a consequence of the popularity and the fast growth of e-business and on-line shopping, an ever growing number of product images are uploaded, viewed, and shared. These images can be organized based on tags, captions, or text descriptions if such information is given. However, many images are created or uploaded without any textual annotation. Indexing and searching these images would require laborious manual tagging, which could be impractical if the collection is very large, motivating content-based retrieval approaches [4][14].

Product packaging images or promotional posters generally have rich text containing category information such as the brands, ingredients, pricing information, etc. which is difficult to extract using standard content-based image retrieval methods. In this paper, we propose an innovative photo OCR-based system for e-business image management. By indexing each image with words recognized from the image automatically, retrieval can be achieved using simple keyword searching.

Conventional OCR methods do not work well on these types of e-business images as such methods were typically designed for scanned documents which generally have structured layouts, horizontal text-lines, clear character blocks and uniform text patterns. As a result, photo OCR techniques are required [1][12]. However, most photo OCR methods were developed for text recognition in generic scenes, and have not been specifically designed for e-business image management. Scene text recognition usually addresses challenges such as low resolution, blurry images, strong lighting and low contrast, etc. E-business images are usually professionally produced and have good image quality in general but they present other challenges for text recognition, such as the cluttered background, various text sizes, location and orientation. Moreover, text in e-business images usually features extremely diverse fonts, and in fact characters are often artistically modified in shape, colour and arrangement.

To address these challenges in e-business images, we take advantage of the combinatorial power of deep neural networks and maximally stable extremal regions (MSERs) processing. Text regions are identified with a convolutional neural network (CNN) model. Letter regions are then extracted as MSERs and grouped as words by pair grouping within the text regions. The word regions are verified by a word/non-word CNN model. A word region growing technique is proposed to complete partially detected word regions. Word regions are then converted to machine coded transcription by a CNN-trained character recognizer. At query time, n-gram vectors are computed for the text query and the transcription. The cosine distance is calculated for searching a pair match between the query and transcription of a detected region and any image that contains a match will be returned from the system.

Several datasets have been published for scene text recognition. However, testing on these datasets could not fully evaluate our proposed system given the specific characteristics of e-business images. Therefore, we prepared a novel dataset[1] specfically designed to reflect the type of text typically encountered in e-business images. The dataset contains 500 e-business images that are categorized into 13 categories. Words in images are annotated with bounding boxes, which provides more than 4000 annotations and approx. 2500 unique words in total.

The rest of the paper is organized as follows: in section 2 we review techniques of text recognition for images. Section 3 details our proposed system. In section 4.1, we describe our dataset and present the challenges of text recognition for e-business images. In section 4.2, we compare our system with two other state-

[1] The dataset is available for download from https://github.com/jiang-public/mmm2018

of-the-art approaches for text recognition. Finally, section 5 ends the paper with some concluding remarks.

2 Related Work

Converting printed or handwritten text in images to machine encoded text has a long history and is usually referred to as optical character recognition (OCR). Conventional OCR methods are designed primarily for black-white scanned documents and typically rely on brittle techniques such as binarization [1][21]. Strong assumptions are usually made in the processing such as the presence of horizontal text lines and that characters have the same size, therefore these methods perform poorly on general images.

Recently, many methods have been proposed for recognizing text in images of natural scenes [12]. Such approaches are usually referred to as PhotoOCR [1]. Some of these methods address sub-tasks such as text detection [8][13] or text recognition [20], while others attempt to combine both to have an end-to-end solution [10][19]. In text recognition, methods typically assume perfect text localization has been achieved and words are “cut-out”. However, in real-world applications, the assumption that text is localized 100% accurately is usually invalid. Text in natural images may exhibit significant diversity of text patterns in highly complicated backgrounds. Therefore, much effort has been devoted to the task of text detection.

Depending on the processing flow for text detection, methods can generally be categorized into two groups, connected component-based methods, and sliding window-based methods. The stroke width transform (SWT) [5] and maximally stable extremal regions (MSERs) [6][16] are two common connected component-based methods. MSER-based detectors exploit the fact that characters normally have strong contrast with the background to allow for easy reading [3][19]. SWT uses the assumption that characters are regions of similar stroke width [10][17]. The same assumption is also exploited in the FASTex detector [2] where text fragments are extracted by local thresholding properties of stroke-specific keypoints. There are also some methods leveraging the benefits of both SWT and MSERs that attempt to achieve improved text localization performance [15][18].

Sliding window based methods process text as a standard task of object detection and recognition [1][11]. Wang et al. [22] detected characters with Random Ferns and grouped them into words with a pictorial structure model. Jaderberg et al. [10] generated candidate word bounding boxes with Edge Box proposals and pruned them with a Random Forest classifier. Each word is then recognized with a CNN model trained on a large synthetic data set [9].

However, no methods have yet achieved sufficient accuracy for practical end-to-end applications. Although promising results have been achieved for scene text recognition [12], the datasets evaluated are not fully realistic – the text patterns are simple, the word orientations are only horizontal, they occupy a significant part of the image, and there is no perspective distortion or significant noise [19]. Therefore, these methods are not expected to perform well on the real

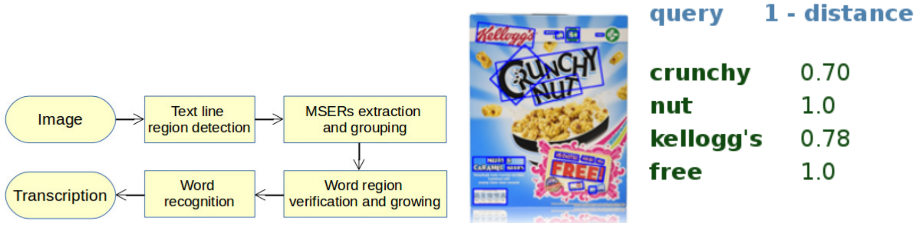


Image
Text line region detection
MSERs extraction and grouping
Word region verification and growing
Word recognition
Transcription
CRUNCHY NUT
FREE!

query	1 - distance
crunchy	0.70
nut	1.0
kellogg's	0.78
free	1.0

(a) Processing pipeline for indexing (b) Illustration of searching

Fig. 1: The backend of our proposed system

challenges aforementioned from the Internet e-business images, as shown later in our experiments.

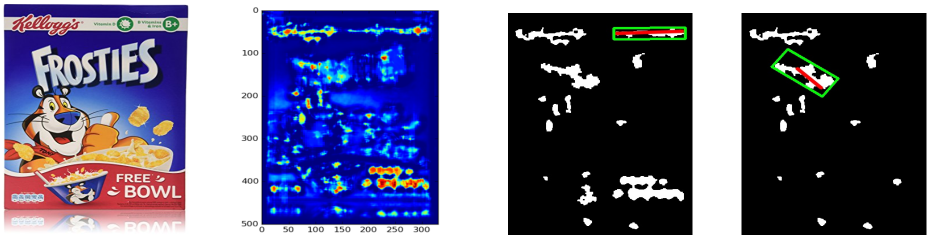
In contrast, our system is designed to tackle those challenges from e-business images. Specifically, (a) an approach of combining the merits from both deep neural networks and MSER processing is developed; (b) A region growing technique is proposed to complete partial detected word regions; (c) we use cosine distances with n-grams vectors instead of keywords comparison with edit distances in retrieval to provide tolerance to inevitable transcription errors in text recognition.

3 The Proposed System

The core of our proposed e-business image management system is its photo OCR-based backend which consists of image indexing and searching. Figure 1 (a) depicts the process pipeline of the image indexing stage. During indexing, text embedded in each image is recognized and its transcription is saved in a database. At query time, a text query is given and a matching transcription is searched in the database based on the cosine distance – sample search terms and their distances are illustrated in figure 1 (b). Any images containing matched transcriptions are retrieved and returned to the user.

3.1 Indexing

Text-line region detection Our system detects text-line regions based on a pretrained CNN model from Oxford [11] and Hough line detection. A text saliency map, shown in figure 2 (b), is generated by evaluating the text/background CNN model, which is trained on cropped 24 × 24 pixel case-insensitive characters, in a sliding window fashion across the image. Hysteresis thresholding ($t_{high} = 0.2$, $t_{low} = -0.5$) is subsequently applied to the saliency map to obtain a binary text/background map. With the binary map, we apply Hough line detection to extract the potential text line regions and remove those sparse text spots which are very likely to be noise detected from the CNN model. Our Hough



(a) Input image (b) Text saliency (c) Text line detection (d) Detected line suppression

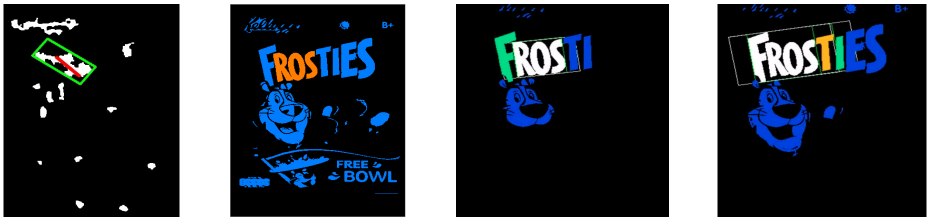
Fig. 2: An example of text-line region detection

line detection is a probabilistic Hough line transform modified as follows. Once a line segment is identified, all pixels from the regions that the line traverses are suppressed and are not used in subsequent line voting. This text line detection is used only to detect the text line regions rather than estimating the baseline of the text, thus curved text line regions can also be identified. Figure 2 shows an example of the detection.

MSERs extraction and grouping The output from text-line region detection is not sufficient for creating accurate word bounding boxes. As the detection for word "kellogg's" in figure 2(b), the detected region is only around the central area of the text. Therefore, we compute MSERs from images for precisely extracting the word regions at a later stage. The MSERs that overlap with a text-line region are pair grouped to form a potential word. Each MSER i searches for its nearest neighbour MSER j. If the normalized distance between i and j is small enough, MSER i and j are considered as a pair of characters from the same word. Given a small evaluation image set, we empirically choose this distance as 1.15.

Word region verification To remove false detections, a word/non-word CNN classifier is trained with Jaderberg's synthetic data [9]. We change the last fully connected layer of Jaderberg's network [9] to have a binary word/non-word output. Each potential word region is classified and empirically those with probability higher than 0.98 are retained.

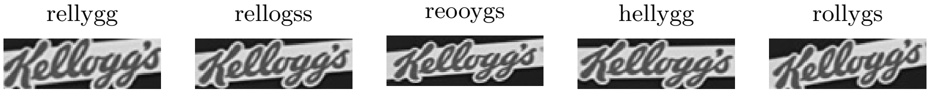
Word region growing A group of retained MSERs may only present a part of a word due to inaccuracies in detection. We propose a word growing technique to address this. As shown in figure 3, for a retained word region we extend the width of the word region from both sides, and search candidate MSERs from all MSERs extracted in the entire image. If any MSER has more than 30% region overlap with the potential word, this MSER becomes a candidate for being attached to the word region. However, among all candidates only the one with the highest overlap ratio is chained to the potential word in any given search. We repeat this search procedure on both left and right directions and grow the potential word continuously until no more candidate MSERs can be found.



(a) A detected text-line region (b) A retained region (c) Search to the left (d) Search to the right

Fig. 3: An example of word region growing. White components in (c) and (d) are MSERS confirmed to be part of a word, green components are candidates being searched, and the orange component in (d) was the winning MSER attached to a word region in the final search

Word recognition After word growing, the word regions are ready for recognition and Jaderberg's character sequence encoding model [9] is applied. We find that the model is sensitive to the padding around the text in the word region. A slight difference in padding could result in different character sequences, especially for cursive characters. Therefore, for each word region we expand the region with 6 and 12 pixel paddings respectively, and rotate the 6 pixels padding region through +5 and −5 degrees such that the word will not be partially cut out because of the rotation. Together with the originally detected region, 5 region patches are extracted and resized to size 32 × 100 individually for translation by the character sequence encoding model as shown in figure 4. Transcriptions from all detected word regions in an image are saved in a database and this image is then indexed.

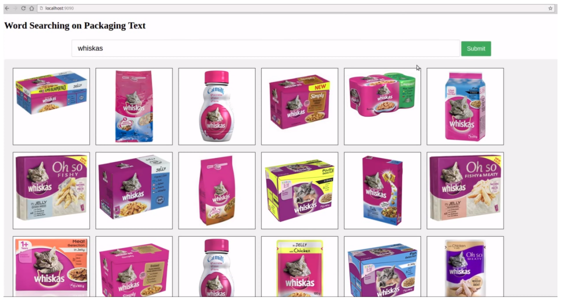


(a) original (b) 6 pixels (c) 12 pixels (d) +5 degrees (e) -5 degrees

Fig. 4: Region patches of "Kellogg's" for word recognition. The top row text is the output from the recognition model in each case

3.2 Searching

The transcription from the character sequence encoding model may not always correspond to meaningful words as illustrated in figure 4. Therefore, the cosine distance with n-gram features is used for string comparison. The n in n-gram is chosen to be 1 to 4. A pre-prepared n-gram look-up table [9] is used to generate a $10K$ dimensional n-gram histogram vector. It should be noted that training a CNN model to convert region patches to n-grams feature vectors directly as described in Jaderberg's paper [9] is inappropriate for our system because this would require our queries to be pictures as well.



Word Searching on Packaging Text

whiskas

Fig. 5: The frontend of the image management system

Transcription from each detected word is compared to the query for matching based on the cosine distance. 5 cosine distances of 5 patches of a detected word are computed. We take the average of these 5 distances to be the distance of the detected word to the query. We iterate through all detected words in an image and then any image containing a matched word is retrieved.

3.3 Frontend

Our proposed system is a web-based management system. Images in a database are pre-indexed asynchronously. When a user submits a query, the system retrieves all the images containing the query word and returns them back to the users through the interface as shown in figure 5.

4 Evaluation

4.1 Dataset

To evaluate our system, a novel dataset is prepared. Comparing to the other public datasets for text recognition evaluation, our dataset is much closer to e-business applications and reveals the challenges of text recognition for e-business images as shown by the examples in figure 6. The dataset contains 500 product packaging and promotional images that are categorized into 13 categories which are artworks, beauty, biscuits, books, boy tops, cereal and porridge, grocery, health and personal care, homecare cleaning, pet supplies, sweets and chocolate, toys, and vhs. Words in images are annotated manually with bounding boxes as the ground truth. One or two character words including numbers are deemed unreadable and not annotated. Words with repeat patterns and unreadable small words are also ignored. This gives us in total $4,123$ annotations and $2,579$ unique words in the dataset.

For detailed evaluation, 390 item queries and 110 document queries were prepared. The item queries includes 6 query sets with 65 words in each set respectively. The "small" words are defined as non-cursive horizontal words with character size around 12×12 pixels, which are usually used for ingredient information, terms and conditions, etc. The "middle" words are those characters

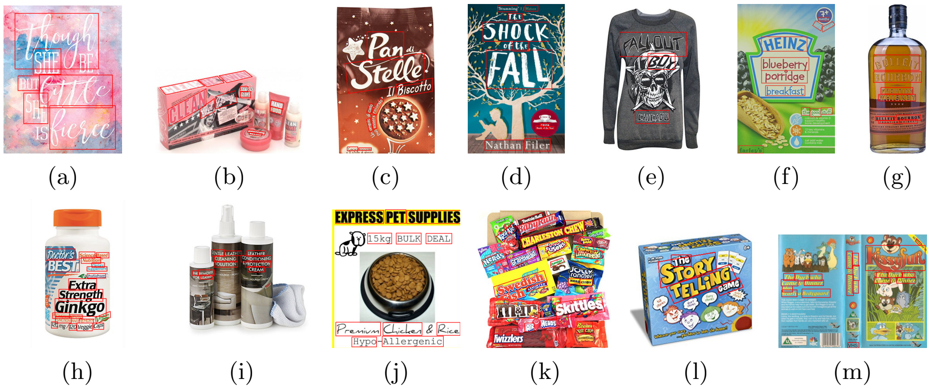


Fig. 6: Examples from the evaluation dataset. (a) *artworks*, (b) *beauty*, (c) *biscuits*, (d) *books*, (e) *boy tops*, (f) *cereal and porridge*, (g) *grocery*, (h) *health and personal care*, (i) *homecare cleaning*, (j) *pet supplies*, (k) *sweets and chocolate*, (l) *toys*, (m) *vhs*

around 24 × 24 pixels, which usually give supplementary information about products such as "all skin types", "chocolate", "international", etc. Characters in "big" words are bigger than 48 × 48 pixels which are generally used for brand names, product names, book names, etc. Characters in the "orientation" set are not horizontally aligned. The words in the "cursive" set are words with handwritten script fonts. The "art" set is the most challenging set, and words have strong artistic decoration, e.g. unequal character size, gradually changed font colour or almost transparent characters.

The 110 document queries were prepared for image retrieval evaluation. The query words were randomly selected from the annotated dataset. Given a query word, all images containing the word are considered relevant. In this evaluation, even if the query word appears in multiple locations in a relevant image, having only one location detected and read out correctly is sufficient for the target application.

4.2 Comparison with other approaches

Two other state-of-the-art scene text recognition approaches are implemented for comparison, which detect and convert word regions to machine-coded transcription. The comparison shares the same searching procedure.

The first approach is Chen's robust text detection in natural images [3]. Chen's approach applies both MSERs and stroke width in character detection and uses the conventional OCR engine Tesseract [21] for word recognition, which provides a baseline for comparison. We adapted Saburo Okita's implementation of the approach [2] and kept most parameters as default except the minimum and

[2] https://github.com/subokita/Robust-Text-Detection

maximum MSER size which are set to 10 and 10000 pixels to keep most MSERs for further processing.

The second approach is Neumann's lexicon-free method [19], which has been implemented in OpenCV [3]. Neumann's approach drops the stability requirement of MSERs and selects suitable extremal regions (ERs) by a sequential classifier trained for character detection. Gomez's method [7] is used for grouping arbitrary oriented text, and characters are recognized by a nearest-neighbour classifier trained with features of chain-code bitmaps such that its word recognition is dictionary free.

4.3 Results

Item detection In item detection, the goal is to localize the query word in an image. Taking each annotated word in the dataset as a query, we obtain an overall detection performance. Figure 7 shows the detection recall change by increasing the detected/non-detected threshold to the overlapping ratio α, where α and the recall R_{det} are defined as in equation (1). G_i stands for the ground truth bounding box of a word and D_i is the bounding box of a detected item. For the many-to-one case where many items are detected within a true word bounding box, we only count this once in the correctly detected items calculation.

$$\alpha = \frac{Area(G_i \cap D_i)}{Area(G_i)} \ , \qquad R_{det} = \frac{\text{No. correctly detected items}}{\text{No. relevant items in dataset}} \tag{1}$$

As shown in Figure 7, our proposed method performs better when 50% overlap is applied, which is the general criterion in object detection. Our method is penalized when high overlap ratio is required, however this could be compensated in recognition by applying the n-gram based matching. We also run our system and the comparison approaches on the item queries with a 50% overlap criterion. As shown in table 1, our recall performance outperforms the other two approaches in most cases except for the orientation and the art query sets. This is probably because stroke width is rotation invariant and the distance transform is less affected by the colour changes within characters, while the character spotting CNN model deployed in our system would require more orientated and art character samples during training. It can also be seen that our text region detection is scale sensitive. The system performs well when the character size is similar to the training samples. Small characters tend to be regarded as noise in our detection, and the word region growing technique compensates more for the big characters than for small ones. Both Chen's and Neumann's methods encounter difficulties in detecting cursive characters. Chen's method assumes characters have low variation in stroke width, and Neumman's method was designed mainly for block print characters. In contrast, our proposed method can still detect a certain amount of cursive words. The precision table shows that all methods suffer from a high false detection rate, with Neumann's method performing a little better than others overall.

[3] http://docs.opencv.org/3.0-beta/modules/text/doc/text.html

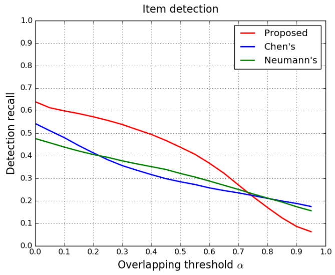


Fig. 7: Item detection recall change by increasing α

Recall	small	middle	big	orientation	cursive	art
Proposed	**0.369**	**0.692**	**0.615**	0.338	**0.492**	0.338
Chen's	0.231	0.169	0.385	**0.430**	0.277	**0.385**
Neumann's	0.354	0.415	0.154	0.292	0.046	0.308

Precision	small	middle	big	orientation	cursive	art
Proposed	0.042	0.070	0.086	0.046	0.090	0.053
Chen's	0.025	0.019	0.038	0.039	0.028	0.039
Neumann's	0.105	0.097	0.055	0.102	0.019	0.113

Table 1: Item detection recall & precision on 6 item query sets

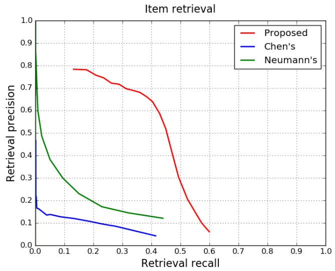


Fig. 8: Item retrieval precision-recall curve

Recall	small	middle	big	orientation	cursive	art
Proposed	**0.385**	**0.662**	**0.585**	**0.292**	**0.262**	**0.354**
Chen's	0.015	0.015	0.031	0.0	0.0	0.015
Neumann's	0.0	0.0	0.0	0.0	0.0	0.0

Precision	small	middle	big	orientation	cursive	art
Proposed	0.641	0.878	0.792	0.559	0.85	0.622
Chen's	1.0	0.167	0.222	0.0	0.0	1.0
Neumann's	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: Item retrieval recall & precision on 6 item query sets

Item retrieval Detected regions are converted to machine-coded transcription. An item is recognized and retrieved if its averaged cosine distance to a query is less than a threshold β. We fix the overlap ratio as 0.1 in detection such that all three methods can keep most regions for query matching. The retrieval performance will thus be strongly impacted by performance of the text recognition. As with item detection, we obtain an overall retrieval performance by iterating through all annotated words in the dataset. Figure 8 plots the Precision-Recall curves of the three methods. The item retrieval recall R_{ret} and precision P_{ret} are defined similar to the definition in item detection by replacing the detected items with retrieved items.

As shown in the plot, Chen's and Neumann's approaches do not perform well on the dataset. This can also be seen from the results in table 2, running the evaluation on the 6 item query sets with distance threshold β 0.5. In Chen's approach, Tesseract was used for text recognition, which was originally designed for scanned document OCR. It assumes print letters on white background and fonts are general sizes used in documents, which is not the case in the proposed dataset. Neumann's approach was mainly designed for block characters in scene text recognition. The character recognition classifier is trained with chain-code features extracted from synthetic black print letters on white background with no distortion, blurring, scaling or rotation introduced, while in our dataset there are a large number of cursive characters and various text patterns.

Image retrieval To retrieve images, each returned image contains at least one detected word matching the input query. Although the retrieval performance of the system has been demonstrated in item retrieval evaluation, we test our system with the 110 document queries and obtain a 0.394 mean average precision. Figure 9 shows retrieved images for an example query "free". It can be seen that, even though some detection is not perfect, our system is still capable of flagging a match owing to the n-grams based string comparison.

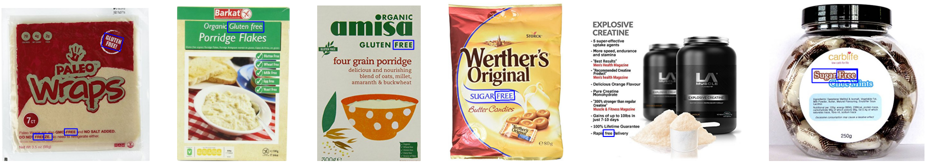


Fig. 9: Correctly retrieved image examples for the query "free"

5 Conclusion

In this paper, we address the challenge text recognition for e-business image management. A method combining neural networks and MSERs for image indexing is proposed. A novel dataset is prepared specially for text recognition evaluation in e-business images, and our system outperforms other two state-of-the-art scene text recognition approaches. In future work, the system will be integrated with other content based image retrieval methods to make image management even easier.

Acknowledgements This research has been supported by Science Foundation Ireland under grant number SFI/12/RC/2289 which is co-funded by the European Regional Development Fund.

References

1. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: Photoocr: Reading text in uncontrolled conditions. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. pp. 785–792. ICCV '13, IEEE Computer Society, Washington, DC, USA (2013)
2. Bušta, M., Neumann, L., Matas, J.: Fastext: Efficient unconstrained scene text detector. In: 2015 IEEE International Conference on Computer Vision (ICCV 2015). pp. 1206–1214. IEEE, California, US (December 2015)
3. Chen, H., Tsai, S.S., Schroth, G., Chen, D.M., Grzeszczuk, R., Girod, B.: Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In: 2011 IEEE International Conference on Image Processing. Brussels (sep 2011)
4. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Information Retrieval 11(2), 77–107 (2008)
5. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: CVPR. pp. 2963–2970. IEEE (2010)

6. Forssén, P.E.: Maximally stable colour regions for recognition and matching. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, IEEE, Minneapolis, USA (June 2007)
7. Gómez, L., Karatzas, D.: Multi-script text extraction from natural scenes. In: Proceedings of the 2013 12th International Conference on Document Analysis and Recognition. pp. 467–471. ICDAR '13, IEEE Computer Society, Washington, DC, USA (2013)
8. He, T., Huang, W., Qiao, Y., Yao, J.: Text-attentional convolutional neural network for scene text detection. Trans. Img. Proc. 25(6), 2529–2541 (Jun 2016)
9. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. In: NIPS Deep Learning Workshop (2014)
10. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. Int. J. Comput. Vision 116(1), 1–20 (Jan 2016)
11. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV. pp. 512–528. Springer International Publishing, Cham (2014)
12. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S.K., Bagdanov, A.D., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Shafait, F., Uchida, S., Valveny, E.: Icdar 2015 competition on robust reading. In: ICDAR. pp. 1156–1160. IEEE Computer Society (2015), relocated from Tunis, Tunisia
13. Koo, H.I., Kim, D.H.: Scene text detection via connected component clustering and nontext filtering. IEEE Trans. Image Processing 22(6), 2296–2305 (2013)
14. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 2(1), 1–19 (Feb 2006)
15. Li, Y., Lu, H.: Scene text detection via stroke width. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 681–684 (Nov 2012)
16. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British Machine Vision Conference. pp. 36.1–36.10. BMVA Press (2002), doi:10.5244/C.16.36
17. Neumann, L., Matas, J.: Scene text localization and recognition with oriented stroke detection. In: 2013 IEEE International Conference on Computer Vision (ICCV 2013). pp. 97–104. IEEE, California, US (December 2013)
18. Neumann, L., Matas, J.: Efficient scene text localization and recognition with local character refinement. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 746–750. IEEE, California, US (Aug 2015)
19. Neumann, L., Matas, J.: Real-time lexicon-free scene text localization and recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 38(9), 1872–1885 (Sept 2016)
20. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. CoRR abs/1507.05717 (2015)
21. Smith, R.: An overview of the tesseract ocr engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02. pp. 629–633. ICDAR '07, IEEE Computer Society, Washington, DC, USA (2007)
22. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Vision. pp. 1457–1464 (Nov 2011)