A Real-Time Scene Text to Speech System

Lukáš Neumann and Jiří Matas

Centre for Machine Perception, Department of Cybernetics
Czech Technical University, Prague, Czech Republic
{neumalu1,matas}@cmp.felk.cvut.cz
http://textspotter.felk.cvut.cz/

Abstract. An end-to-end real-time scene text localization and recognition method is demonstrated. The method localizes textual content in images, a video or a webcam stream, performs character recognition (OCR) and "reads" it out loud using a text-to-speech engine. The method has been recently published, achieves state-of-the-art results on public datasets and is able to recognize different fonts and scripts including non-latin ones.

The real-time performance is achieved by posing the character detection problem as an efficient sequential selection from the set of Extremal Regions (ERs) which has a linear computation complexity in the number of pixels in the image. Robustness to blur, noise and illumination and color variations is also demonstrated. Finally, we show effects of various control parameters.

1 Introduction

Text localization and recognition in real-world (scene) images is an open problem which has been receiving significant attention by the computer vision community. Localizing text in an image is potentially a computationally very expensive task as generally any of the 2^N subsets can correspond to text (where N is the number of pixels). Numerous methods have been recently published [1,2,3,4,5,6], however they only focus on text localization and their performance is not real-time.

In this demonstration, we present a recently published end-to-end real-time[1] text localization and recognition method [7], which achieves state-of-the-art results on standard datasets (see Figure 2). The real-time performance is achieved by posing the character detection problem as an efficient sequential selection from the set of Extremal Regions (ERs). The ER detector is robust against blur, low contrast and illumination, color and texture variation. Its complexity is $O(2pN)$, where p denotes number of channels (projections) used. The method is able to recognize different fonts and scripts including non-latin ones (see Figure 3).

[1] We consider a text recognition method real-time if the processing time is comparable with the time it would take a human to read the text.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 619–622, 2012.

2 Method Description

In the first stage of the classification, the probability of each ER being a character is estimated using features calculated with $O(1)$ complexity and only ERs with locally maximal probability are selected for the second stage, where the classification is improved using more computationally expensive features. A highly efficient exhaustive search with feedback loops [8] is then applied to group ERs into words, select the most probable character segmentation and perform character recognition (OCR).

Each frame of the webcam video stream is processed independently and text from subsequent frames is aggregated. If the same word is recognized in the same place of the image (with some tolerance) in sufficient number of subsequent frames, it is considered as detected and it is passed to the speech engine (see Figure 1).

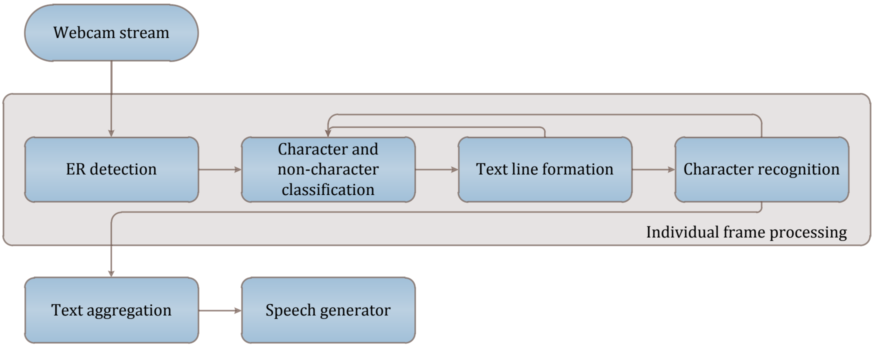
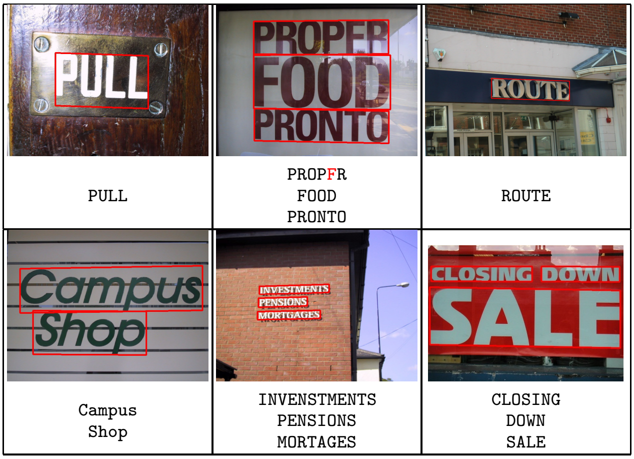


Fig. 1. Stages of the demonstrated method. Each frame is processed independently and text is aggregated for subsequent frames.

3 Application Interface

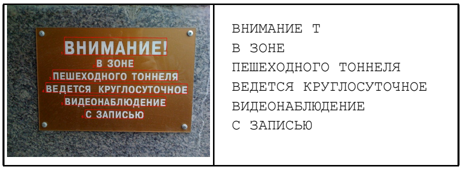
The demo application interface consists of two independent dialogs: a control console and an output window (see Figure 4). In the control console, a user can select source to be processed (a directory, a single file or a webcam stream), set control parameters and select output mode of the method. The control parameters demonstrate different configurations of the method, such as the trade-off between speed and quality with increasing number of projections. Each output mode shows different stage of the processing and user is thus able to see the role of each module incorporated in the demonstrated method (see Figure 5).

Current work includes porting the method on mobile devices and improving visualization capabilities.



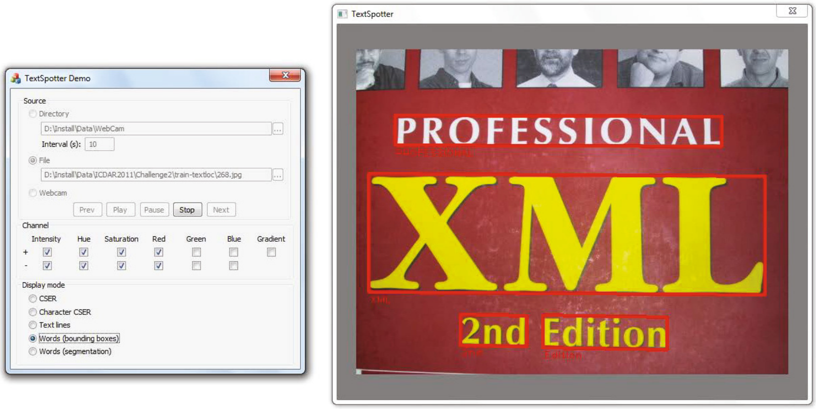
```
PULL

PROPFR
FOOD
PRONTO

ROUTE

Campus
Shop

INVENSTMENTS
PENSIONS
MORTAGES

CLOSING
DOWN
SALE
```

Fig. 2. Text localization and recognition examples on the ICDAR 2011 dataset. Notice the robustness against reflections and lines passing through the text (bottom-left)



```
ВНИМАНИЕ Т
В ЗОНЕ
ПЕШЕХОДНОГО ТОННЕЛЯ
ВЕДЕТСЯ КРУГЛОСУТОЧНОЕ
ВИДЕОНАБЛЮДЕНИЕ
С ЗАПИСЬЮ
```

Fig. 3. Text localization and recognition output example on non-latin script



TextSpotter Demo

Source

○ Directory

D:\Install\Data\WebCam

Interval (s): 10

● File

D:\Install\Data\ICDAR2011\Challenge2\train-textloc\268.jpg

○ Webcam

Channel

	Intensity	Hue	Saturation	Red	Green	Blue	Gradient
+	☑	☑	☑	☑	☐	☐	☐
-	☑	☑	☑	☑	☐	☐	

Display mode

○ CSER
○ Character CSER
○ Text lines
● Words (bounding boxes)
○ Words (segmentation)

Fig. 4. Application interface. Control console (left) and output window (right)

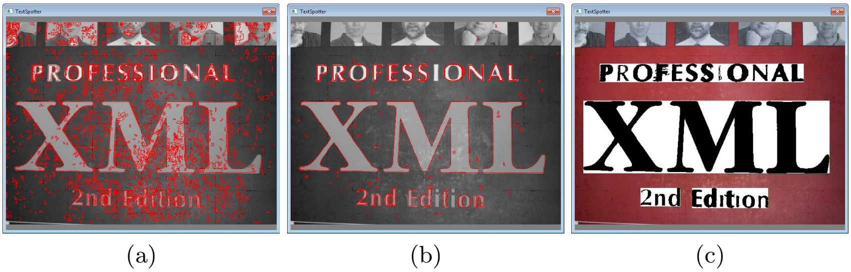


Fig. 5. Application output modes. Class-Specific Extremal Regions (a). Character Extremal Regions (b). Resulting Segmentation (c).

4 Conclusions

An end-to-end text localization and recognition system will be presented. The method is real-time and it achieves state-of-the-art results on standard datasets.

Acknowledgments. The authors were supported by the Czech Science Foundation Project GACR P103/12/G084 and by the CTU Prague student project SGS11/125/OHK3/2T/13.

References

1. Jung-Jin, L., Lee, P.H., Lee, S.W., Yuille, A., Koch, C.: Adaboost for text detection in natural scene. In: ICDAR 2011, pp. 429–434 (2011)
2. Chen, X., Yuille, A.L.: Detecting and reading text in natural scenes. In: CVPR, vol. 2, pp. 366–373 (2004)
3. Lienhart, R., Wernicke, A.: Localizing and segmenting text in images and videos. Circuits and Systems for Video Technology 12, 256–268 (2002)
4. Pan, Y.F., Hou, X., Liu, C.L.: Text localization in natural scene images based on conditional random field. In: ICDAR 2009, pp. 6–10. IEEE Computer Society (2009)
5. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: CVPR 2010, pp. 2963–2970 (2010)
6. Zhang, J., Kasturi, R.: Character Energy and Link Energy-Based Text Extraction in Scene Images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 308–320. Springer, Heidelberg (2011)
7. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: CVPR 2012 (to appear, 2012)
8. Neumann, L., Matas, J.: Text localization in real-world images using efficiently pruned exhaustive search. In: ICDAR 2011, pp. 687–691 (2011)