TEXT DETECTION IN NATURAL SCENES WITH MULTILINGUAL TEXT

Mikhail Zarechensky, Scientific supervisor: Ph.D. Natalia Vassilieva
Department of Analytical Information Systems, St. Petersburg State University
zarechenskij@gmail.com

Abstract

Detecting text in natural scenes is an important prerequisite for further text recognition and other image analysis tasks. Most of text detection methods for scene images usually use a priori knowledge of language to detect text. As a rule such algorithms are evaluated on datasets which contain scenes only with text in English. This paper discusses known text detection algorithms and investigates them for invariance to the language.

1. Introduction

Recent advances in digital technology allow to take pictures from a large number of mobile devices. As a result, the number of photos taken by users is increasing every day. At the same time, we often have no annotations for images except those made by the device. Text in images provides important information about semantics of the image. Annotated images can be used in various applications, such as content-based image retrieval, automatic navigation, automatic translation. It is often the case that a language of a text in an image is not known in advance, or a single image contains text areas with text in different languages. How to effectively detect and recognize text in scene images is an actual research question. Text detection is an important prerequisite for further text recognition. In this paper we explore the problem of text detection.

In this paper, we discuss several known text detection algorithms and investigate them for invariance to a language. Quality of the text detection algorithm greatly depends on the shooting conditions and noise on the image, but in this paper we focus on the problem of language invariance of the algorithms in good conditions. First, we distinguish the main common steps of these algorithms. Second, we provide a theoretical estimation of language invariance for every step of the algorithms. Third, we perform experiments with two algorithms on different datasets to confirm theoretical result.

2. Related work

In order to recognize text in an image, it first has to be robustly detected. Unlike text detection for document images, text detection for scenes is still a challenging task due to the large variety of text appearance in images. Text in scenes can have different variations of the font style, size, distortion; it can have different contrast due to different lighting conditions. The whole image can also vary greatly. We should take into account low resolution, low contrast, heterogeneous background. Such variety gives rise to various approaches to text detection.

Existing methods for scene text detection can be broadly categorized into three groups: texture-based methods, region-based methods and hybrid methods.

The basic difference between these techniques of text detection methods is described in [12]. In this paper we consider only connected components based methods. According to the results of the competition at the ICDAR 2013, this approach proved itself to be more effective comparing to others. We picked methods proposed by Yin et al.[11], Gomez et al.[5] and Chen et al.[2] for further consideration. These algorithms have good results on the ICDAR datasets and use different approaches to detect text.

3. Overview of text detection algorithms

The algorithm proposed by Yin et al.[11] was presented at ICDAR 2013 and got the first place in «Multi-script Robust Reading Competition in ICDAR 2013»[6]. It uses an approach based on the MSER algorithm to find character glyphs.

A disadvantage of the MSER[7] is that it detects a lot of false positives -- regions that do not contain characters. To solve this problem, the algorithm by Yin et al.[11] additionally performs parent-children elimination for the MSER tree. It improves accuracy of finding character regions. The main idea is to eliminate regions with very small or very big aspect ratio.

The next step of the algorithm is to group characters in order to construct text candidates. Character candidates are clustered into text candidates by the single-link clustering algorithm. The parameters depend on the following features: spatial distance, a differences between width and height, top and bottom alignments, color difference, stroke width difference.

At the final step text candidates are labeled by a classifier as text or non-text areas. The following features are used to train the classifier: smoothness, the average stroke width, stroke width variation, height, width, and aspect ratio.

As described in [12] for algorithms presented in [2], [5], text features that

are using to filter out non-characters are similar to features discussed above.

4. Analysis of the main steps of the algorithms

In this section we discuss the main steps of the methods and provide a theoretical estimation of their language invariance.

Character candidates extraction

As presented above, for the region decomposition it is common to use the MSER algorithm. The MSER algorithm depends only on the intensity of the image. Since the text in the image tends to have equal intensity, at least in each symbol, the result of the algorithm is independent of language.

We can conclude that the MSER algorithm is equally applicable for region decomposition as for images, containing only one language and for images with multilingual text. As the Canny edge detector not depend on a language, it follows that the modified MSER, proposed by Chen et al. is also invariance to a language

Filtering of regions

Let us review every feature used for region filtering.

- Aspect ratio

Most letters of English language have aspect ratio being close to 1, so this feature might be useful to filter out false character candidates. To cope with elongated letters such as 'i' or 'l', a threshold should be small enough. On the one hand, this feature can be used for many languages because even if a letter has a very small aspect ratio and is filtered out, the absence of this letter will not affect the grouping of whole word at the grouping stage.

On the other hand, when an entire word is not split into the characters it might cause difficulties in text detection. There are languages in which every word is continuously connected. For instance, Hindi, in which all words are linked by continuous line. In this case, rational use of this feature is difficult, because words might be very long.

Thus this feature has limitations and may not be used for all languages.

- Region height

Irrespective of language, height of the characters in one word are always about the same. Therefore, this type of filter is invariant to the language.

- Number of holes

Number of holes in the English characters and in the hieroglyphs might be different. Therefore, this feature requires an additional configuration for different languages.

- Stroke width

This feature is very important as it is shown in the work Epshtein et al.[4]. However, the proposed implementation has a limitation for the elements that have non-parallel edges. This feature of the implementation is essential for such languages as Arabic. Also the style of writing in Arabic language tends to have more variation in the stroke width, thus to achieve maximum efficiency, this feature must be configured for different languages separately.

5. Empirical analysis

To confirm the theoretical estimations provided in the previous section we will perform a series of experiments for the two methods described in the section 2.

Description of the experiments

For the experiments was selected the algorithm proposed by Yin et al.[11].

For evaluation we used a similar approach and the same quality measures as in the evaluation scheme of ICDAR 2013 competition. The following quality measures are used: precision, recall and f-measure.

Description of test data

In the first group of tests we run the selected algorithm on the following datasets: MSRA-TD500, ICDAR 2011, ICDAR 2013. ICDAR 2011 dataset contains images with text in English only. MSRA-TD500 dataset contains images with text in English and Chinese. And ICDAR 2013 dataset contains images with multilingual text including Indo-Aryan languages and Chinese writing.

Also we created a new synthetic dataset which contains images with multilingual text. Dataset split by languages with prevailing text features described above: stroke width, alignment, aspect ratio, etc. Each group corresponding to the one language contains 100 images. Specific image contains only one word and simple background to focus only on the text features.

Example from synthetic dataset with text in Urdu

Analysis of the experimental results

The results of every test are presented in the following table.

Dataset	Recall	Precision	F-measure
ICDAR 2011	0.68	0.86	0.76
ICDAR 2013	0.42	0.64	0.51
MSRA-TD500	0.21	0.52	0.30

The results of the algorithm on the synthetic dataset are presented in the following table.

Features	Language	Recall	Precise	F-measure
	Chinese	0.64	0.72	0.68
	English	0.57	0.73	0.64
Stroke width	Sinhala	0.48	0.69	0.57
Alignment	Maldivian	0.33	0.69	0.44
Alignment, Stroke width	Urdu	0.23	0.63	0.33
Aspect ratio	Bengali	0.10	0.82	0.18
Aspect ratio	Marathi	0.06	0.91	0.11

Based on the experimental results one may see that the difference between the result on the ICDAR 2011 dataset which contains only images with English text, and all others, is quite big.

From the experimental results on the synthetic dataset we can conclude that quality of some text detection algorithms could strongly depend on such text features as aspect ratio, alignment and stroke width.

6. Conclusion

In this work the following results were obtained.

1. The most efficient text detection algorithms are discussed.
2. The main common steps of text detection algorithms are identified.
3. Every step of text detection algorithms is analyzed analytically for

invariance to a language.

4. Evaluated a series of experiments
5. Created synthetic dataset

During the work it was obtained that the existing set of features may strongly depend on a language. By changing settings of the rules that are used in the algorithms you can improve the text detection results on some pre-defined languages.

As a possible continuation to this work it is planned to implement a complete algorithm that solves the problem of text detection irrespective of a language. The analysis presented in this paper helps to identify problem pieces of the existing algorithms. The created dataset and the experimental results will allow to evaluate better the result of this new algorithm.

References

1. J. Canny. A computational approach to edge detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679–698, 1986.
2. H. Chen, S.S. Tsai, G.Schroth, David M. Chen, R. Grzeszczuk, and B. Girod. Robust text detection in natural images with edge- enhanced maximally stable extremal regions. IEEE International Conference on Image Processing, Sep 2011.
3. A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. IEEE Trans. PAMI, 2003.
4. B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In CVPR, 2010.
5. L. Gomez and D. Karatzas. Multi-script text extraction from natural scenes. ICDAR, 2013.
6. D. Kumar, M.N. Anil Prasad, and A.G. Ramakrishnan. Multi-script robust reading competition in icdar 2013. In ACM Proc. International Workshop on Multilingual OCR, (MOCR 2013), 2013.
7. J. Matas, O. Chum, M. Urban, and T. Pajdl. Robust wide baseline stereo from maximally stable extremal regions. In British Machine Vision Conference, volume 1, pages 384–393, 2002.
8. L. Neumann and J. Matas. Real-time scene text localization and recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2012.
9. N. Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, 9(1):62–66, 1979.
10. I. Zeki Yalniz, Douglas Gray, and R. Manhmatha. Adaptive exploration of text regions in natural scene images. ICDAR, 2013.
11. X.-C. Yin, X. Yin, and K. Huang. Robust text detection in natural scene images. CoRR, abs/1301.2628, 2013.
12. M. Zarechensky. Text detection in natural scenes with multilingual text. SYRCoDIS, 2014.