SNOOPERTEXT: A MULTIRESOLUTION SYSTEM FOR TEXT DETECTION IN COMPLEX VISUAL SCENES

R. Minetto[(1,4)], *N. Thome*[(1)], *M. Cord*[(1)], *J. Fabrizio*[(2)], *B. Marcotegui*[(3)]

(1) UPMC Univ Paris 06, LIP 6, 4 place Jussieu, 75005 Paris, France
(2) LRDE - EPITA, 14-16, rue Voltaire, 94270 Le Kremlin-Bicêtre, France
(3) MINES Paristech, CMM, 35 rue Saint Honoré, 77305 Fontainebleau, France
(4) University of Campinas UNICAMP, Av. Albert Einstein, 1251, Campinas-SP, Brazil

ABSTRACT

Text detection in natural images remains a very challenging task. For instance, in an urban context, the detection is very difficult due to large variations in terms of shape, size, color, orientation, and the image may be blurred or have irregular illumination, *etc*. In this paper, we describe a robust and accurate multiresolution approach to detect and classify text regions in such scenarios. Based on generation/validation paradigm, we first segment images to detect character regions with a multiresolution algorithm able to manage large character size variations. The segmented regions are then filtered out using shape-based classification, and neighboring characters are merged to generate text hypotheses. A validation step computes a region signature based on texture analysis to reject false positives. We evaluate our algorithm in two challenging databases, achieving very good results.

Index Terms— Text detection, multiresolution, image segmentation, machine learning.

1. INTRODUCTION

Text detection is a challenging task in Computer Vision. Many approaches have been proposed, but most of them are dedicated to specific contexts, such as automatic localization of postal addresses on envelopes [1], license plate localization [2], *etc*. For natural scene processing, more generic systems have been recently considered [3, 4, 5]. Efficient text detection should provide useful information for many applications related to scene understanding. However, no standard efficient solution really emerges on urban context.

ICDAR conference organized *robust reading competitions* [1] which goal was text detection that go beyond current O.C.R. performances. Such competitions establish a common benchmark giving a clear understanding of the current state of the art of such algorithms [5]. Two types of systems have been proposed, bottom-up approaches based on character identification followed by a grouping step for text detection, and top-down techniques looking first for text regions and then for characters. Hinnerk Becker [5] (best results to the ICDAR challenge) developed a bottom-up approach that uses an adaptative binarization scheme to extract character regions which are then combined fulfilling some geometrical constraints to create text lines. Alex Chen et al [5] (second best in the ICDAR challenge) developed a top-down approach that makes use of a statistical analysis over text regions to select relevant features for characterizing text. Then, they use a cascade of classifiers trained over the chosen features to select regions candidates. Finally, connected components are extracted over these regions, and analyzed to recover each text word.

Although these approaches are interesting, performances are limited and become even very low in a hard urban visual environment. More comprehensive strategies are necessary in order to miss as few text regions as possible, while offering a good robustness to false positive detection.

In this work, we introduce our system, called SnooperText, which is an hybrid scheme combining bottom-up and top-down strategies. Regarding methodology, our approach relies on the hypothesis generation/validation paradigm. The hypothesis generation is carried out using a bottom-up approach: starting from a character segmentation, classification and grouping allow us to provide text region hypotheses. In order to deal with the strong character size variations in urban context, we propose a new multiresolution strategy. Next, the hypothesis validation module, based on a top-down strategy, corresponds to a new classification and works in a text region level using global descriptors (*i.e.* gradient patterns) over the text regions returned by the hypothesis generation. This information is complementary to the bottom-up stages of the algorithm, and grandly decreases the false positive rate.

The remainder of the paper is organized as follows. Section 2 presents the overall system for text detection. Section 3 and 4 point out the two methodological area of novelty of the paper. Section 5 shows that SnooperText is very competitive being among the state-of-the art systems in the ICDAR dataset. Finally section 6 concludes the paper and proposes directions for future works.

2. SYSTEM OVERVIEW

The whole scheme of SnooperText is shown in figure 1. As previously mentioned, our system generates a set of text hypotheses, and validates them using a complementary strategy. Regarding hypothesis generation our algorithm is composed of three main steps: image segmentation, character classification, and character grouping.

The segmentation step is based on a morphological operator, *toggle mapping*, introduced by Serra [6]. Toggle mapping is a generic operator which maps a function on a set of n functions and is generally used for contrast enhancement, noise reduction and recently for image segmentation [7]. The segmentation with the toggle mapping is done by means of morphological erosions and dilations. The advantage of this approach is that it can efficiently detect image boundaries necessary to recognize each image character.

The segmentation produces a set of homogeneous regions. We now aim at discriminating regions that contain text (characters) from those that do not. To achieve this goal, we use a classification strategy based on the extraction of shape descriptors in each image region. We have selected three families of

Contact: minettor@poleia.lip6.fr. This work was partially supported by FAPESP (Phd grant 07/54201-6), CAPES/COFECUB 592/08 and ANR 07-MDCO-007-03.

[1] Simon M. Lucas. http://algoval.essex.ac.uk:8080/icdar2005/

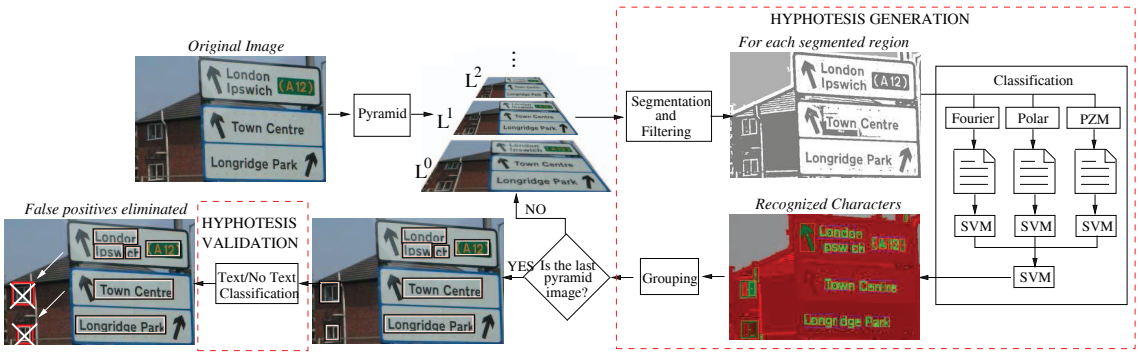


Fig. 1. SnooperText scheme.

descriptors: fourier moments, pseudo zernike moments and a new definition of a polar representation [8]. These descriptors are appealing since they are scale and rotation invariant. Then, a hierarchical SVM classifier [9] is used to discriminate characters from non-character regions. Thus, we train three different classifiers at the fisrt level with each family of descriptors. The final decision is given by merging the previous outputs into a third SVM classifier (Figure 1).

In order to build text hypotheses, we developped a grouping step where all recognized characters are grouped all together with their neighbours to recover the text regions. The conditions to link two characters to each other are those given in [3]. They are based on the distance between the two regions relatively to their height. During this process, isolated text regions (single characters) are eliminated. This aggregation is mandatory to generate words and sentences to integrate as an input in an O.C.R. but it also suppresses a lot of false positive detections. At the end *rectangular windows* are detected in the image. These windows are the input for the hypothesis validation step fully described in section 4.

3. MULTIRESOLUTION SCHEME

We propose a multiresolution segmentation algorithm that extends the approach proposed by Fabrizio *et al.* [7, 8]. This latter method for text segmentation is very efficient, attested by its 2^{nd} rank in the recent ICDAR 2009 Document Image Binarization Contest [10]. However, in complex scenes like urban images where text scale may highly vary, this method is likely to fail, especially in cluttered background with textured areas or local illumination variations. Indeed, the size of the morphological operator intrinsically defines the size of the homogeneous segmented regions. Thus, large text areas with texture are prone to over-segmentation, while small text regions might be missed. To overcome this shortcoming, we propose to apply the algorithm [8] in a multiresolution fashion. Each resolution level is dedicated to detect a given range of text regions scales. At coarser levels we aim at detecting large text areas, and ignoring texture details (high frequencies). In addition, note that the processing is speeded up due to decreasing image size. At finer levels, our goal is to detect smaller regions, analyzing more accurately the local image content.

More formally, let us consider a pyramid image I^l, $l \in \{0; L-1\}$, as defined in the multiresolution framework [11]. At resolution level l, the image is decimated with a ratio of 2, leading to an image size decreased by 4^l (with respect to the initial image). Many solutions can be adopted for relating resolution levels to text region scales. In this paper, we propose a simple yet effective technique, using a fixed size over scales. Thus, at resolution level l, we aim at detecting text regions with size $s_l \in [m_l; m_l + \delta_l]$. In addition, we define a constant c_l corresponding to the overlap between two consecutive scales s_l and s_{l+1} (see figure 2), so that:

$$m_{l+1} = \frac{m_l + \delta_l - c_l}{4} \quad (1)$$

As δ and c are constant over scales, we have: $\delta_l = \delta_0/4^l$ and $c_l = c_0/4^l$, and m_l can be computed as follows:

$$m_l = \frac{m_0 + l(\delta_0 - c_0)}{4^l} \quad (2)$$

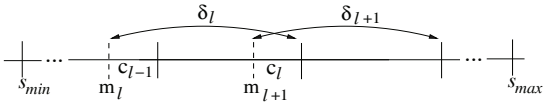


Fig. 2. Region sizes managed at different resolution levels.

Therefore, if we require to detect text regions with size $s \in [s_{min}; s_{max}]$ (where $m_0 = s_{min}$), with a given overlap c_0 in each level and L pyramid images, δ_0 must fulfill : $\delta_0 = (s_{max} - s_{min} + c_0(L-1))/L$.

Practically, our multiresolution algorithm processes as follows. First, we build a set of L pyramid levels, and run the segmentation algorithm of [7] in each downsampled image. At level l, a set of regions r_l^i ($i \in \{1; N_l\}$) are extracted. Only regions whose area s_l^i fall into the bound $[m_l; m_l + \delta_l]$ are considered for the following processing steps, others are ignored. Figure 3 illustrates our multiresolution algorithm. Figure 3(a) shows the image segmentation in the original image resolution ($l = 0$), while Figure 3(b) shows the segmentation in a coarser image resolution ($l = 2$). As we can see in Figure 3(c), when $l = 0$, small text regions are found (yellow windows), when $l = 1, 2$ bigger text regions are found due to the scale intervals computed in equation (2) and thanks to the texture removeal. Figure 3(d) shows the results of the monoresolution approach [8], that fail at detecting the word RIESCOPAM, due to its texture and color.

4. HYPOTHESIS VALIDATION

Since the classification step only analyzes the local image content around each character, false positives occur in complex urban scenes where geometric objects might be confused with characters. Some false positives are shown in figure 4: *e.g.* the bars of the guardrail have a similar shape to a sequence of i's.

(a) Image segmentation when $l = 0$

(b) Image segmentation when $l = 2$

(c) Multiresolution results: detections at scale $l = 0$ in yellow, $l = 1$ in green and $l = 2$ in white

(d) Monoresolution results

Fig. 3. Multiresolution system results, 3 scales and $L = 2$, (a, b and c) and monoresolution results (d). The word "RIESCOPAM" is detected in the multiresolution approach but not in the monoresolution (d).

To deal with these false positives, we apply an hypothesis validation step relying on global image descriptors over the detected windows. These global descriptors are complementary to those used in the hypothesis generation process. For example, in the guardrail case of figure 4, we aim at extracting features encoding periodical patterns that are not present in text regions. In this work, we used the Histograms of Oriented Gradients (HOG) descriptors [12]. HOG descriptors are based on the idea that local object appearance and shape can be well characterized by the distribution of local intensity gradients or edge directions. HOG descriptors proved to reach state of the art performances for object recognition (*e.g.* pedestrian detection [12]), and have been recently used for text detection [4].

To achieve good performances, the HOG extraction in [12] is based on splitting a given window into $N \times M$ cells. This tiling allows to capture spatial information. In addition, in order to be robust to local illumination variations, a contrast normalisation is applied. Groups of $N' \times M'$ adjacent cells are grouped into larger spatial blocks, with an overlap of K cells. Each cell is then normalised with respect to each of its surrounding block. In our experiments, we use $N = M = 4$, $N' = M' = 2$ and $K = 2$. From the extracted feature, we train a SVM classifier to discriminate text from non-text windows. We use a Gaussian chi2 kernel, and perform a cross-validation to optimize its standard deviation σ parameter. Figure 4 illustrates some non-text windows that have been successfully discarded after our hypothesis validation step.

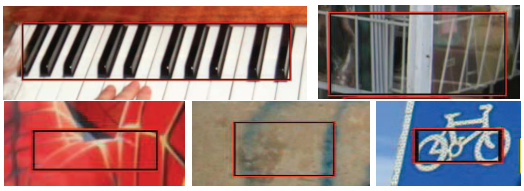


Fig. 4. Hypothesis validation: the windows above were correctly identified as non-text by our validation step.

5. EXPERIMENTS

To validate our system performances, we evaluate the proposed approach on two datasets. First, we run the ICDAR competition[1] to evaluate our approach on a publicly available database. The ICDAR database is composed of 509 real available images (split into training and testing) with complex backgrounds. To evaluate our algorithm performances we use the metrics defined in [5]. The precision and recall were defined as: $p = (\sum_{r_e \in E} m(r_e, T))/|E|$ and $r = (\sum_{r_t \in T} m(r_t, E))/|T|$, where $m(r, R)$ defines the best match for a rectangle r in a set of rectangles R, T and E are the groundtruth sets and estimated rectangles respectively. To combine the precision and recall we use the f measure defined as [5]: $f = 1/(\alpha/p + (1-\alpha)/r)$ where α is a weighting coefficient (set to 0.5).

System	precision	recall	f
SnooperText	**0.63**	0.61	0.61
Hinnerk Becker	0.62	**0.67**	**0.64**
Alex Chen	0.60	0.60	0.60
Ashida	0.55	0.46	0.50
HWDavid	0.44	0.46	0.45
[8]	0.46	0.39	0.43
Wolf	0.30	0.44	0.36
Qiang Zhu	0.33	0.40	0.36
Jisoo Kim	0.22	0.28	0.25
Nobuo Ezaki	0.18	0.36	0.24
Todoran	0.19	0.18	0.18
Full	0.01	0.06	0.07

Table 1. ICDAR performance results.

Table 1 gathers the performance evaluation in the ICDAR database. This evaluation validates the fact that SnooperText is very competitive among state of the art methods in a generic database. Indeed, we rank 2^{nd} regarding f measurement, and 1^{st} regarding precision. In addition, we want to stress that our algorithm was not specifically designed to all images present in this database, *e.g.* images with just one character, hand-writen, *etc*. However, even in these difficult cases, with blurred text regions, bad defined characters, *etc*, our algorithm successfully detects most of the text regions, as illustrated in figure 5.

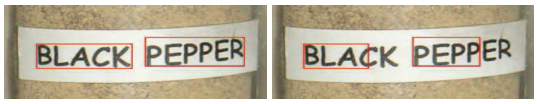
To quantify the gain brought out by our two main contributions (multiresolution algorithm and hypothesis validation step), we also evaluate the performances obtained using only the hypothesis generation block (proposed in [8]). As we can see in table 1, [8] get a precision and recall of 0.46 and 0.39, while our overall system gets a precision and recall of 0.63 and

Fig. 5. Text detection of SnooperText in the ICDAR database.

Fig. 6. Text detection of SnooperText in the iTowns database.

0.61 respectively. This proves the significant performance improvement due to the two area of novelty of this paper. The processing time of our approach depends on the image size and the scene complexity. It takes about 1 minute for a complex image of size 640×480 pixels and about 15,000 regions.

Another crucial aspect is the definition of the metric. Indeed, even using the same database, a change in the evaluation metric can have a significant impact on the performance evaluation. For instance, considering the groundtruth shown in Figure 7(a), if we use the metric of the recent papers [13, 14], to the text detection shown in Figure 7(b), we have precision and recall of 1.0. However, if we use the ICDAR metric we have precision and recall of 0.81. Therefore, to be fair in the comparison and to use a common benchmark, we compare in this paper only approaches that use the ICDAR database and metric.



(a) (b)

Fig. 7. Evaluating different metrics: (a) groundtruth, (b) text detection. Using different metrics to evaluate (b) p and r can vary from 0.81 to 1.0.

We also evaluate SnooperText in a urban scene image database. In line with the iTowns project [2], we manually collect and annotate 100 images of complex urban scenes, with various types of text regions. We make these images and their XML groundtruth annotation publicly available [3]. For example, figures 3 and 6 are two samples of this dataset and illustrate its difficulty (large text size variation, complex background, *etc*). We evaluate our system in this database and again compare performances with respect to the approach [8], as shown in table 2. As we can see, the precision increases in more than 10% with our approach, when the SVM bias b is 0. In this case, some correctly detected text regions are wrongly classified as non-text regions by our classification scheme, leading to a slight drop in the recall result. However, adjusting b so that the recall is 52%, we observe an increase in precision of 5%, supporting the efficiency of our approach.

[2]iTowns ANR project. http://www.itowns.fr

[3]iTowns Dataset. http://www-poleia.lip6.fr/∼minettor/itowns.html

System	precision	recall	f
SnooperText (b = 0)	**0.46**	0.49	**0.48**
SnooperText (b = -0.12)	0.40	**0.52**	0.46
[8]	0.35	**0.52**	0.44

Table 2. iTowns performance results.

6. CONCLUSION

We have proposed a complete system for text detection in complex natural images. Focussing first on character segmentation, filtering and grouping, we generate text region hypotheses. This process is embedded in a multiresolution scheme to handle text regions of various sizes. A validation step exploiting region signature based on texture analysis allows to filter a lot of false positives. We have evaluated our scheme in two databases, achieving very good results. As shown in experiments, our multi-scale approach significantly improves text detection performances with respect to a single-scale approach.

7. REFERENCES

[1] Paul W. Palumbo, Sargur N. Srihari, Jung Soh, Ramalingam Sridhar, and Victor Demjanenko, "Postal address block location in real time," *Computer*, vol. 25, no. 7, pp. 34–42, 1992.

[2] Nicolas Thome, Antoine Vacavant, Lionel Robinault, and Serge Miguet, "A cognitive and video-based approach for multinational license plate recognition," *Machine Vision and Applications*, March 2010.

[3] Thomas Retornaz and Beatriz Marcotegui, "Scene text localization based on the ultimate opening," *ISMM*, vol. 1, pp. 177–188, 2007.

[4] S.M. Hanif, L. Prevost, and P.A. Negri, "A cascade detector for text detection in natural scene images," in *19th ICPR*, Dec. 2008, pp. 1–4.

[5] S.M. Lucas, "Icdar 2005 text locating competition results," in *8th ICDAR*, 2005, pp. 80–84 Vol. 1.

[6] Jean Serra, "Toggle mappings," *From pixels to features*, pp. 61–72, 1989, J.C. Simon (ed.), Elsevier.

[7] Jonathan Fabrizio, Beatriz Marcotegui, and Matthieu Cord, "Text segmentation in natural scenes using toggle-mapping," *IEEE ICIP*, 2009.

[8] Jonathan Fabrizio, Matthieu Cord, and Beatriz Marcotegui, "Text extraction from street level images," *City Models, Roads and Traffic (CMRT)*, 2009.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] Basilis Gatos, Konstantinos Ntirogiannis, and Ioannis Pratikakis, "Document image binarization contest (dibco)," in *ICDAR*, 2009, pp. 1375–1382.

[11] Jean-Michel Jolion and Azriel Rosenfeld, *A Pyramid Framework for Early Vision: Multiresolutional Computer Vision*, 1994.

[12] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*. 2005, pp. 886–893, IEEE Computer Society.

[13] K. C. Kim, H. R. Byun, Y. J. Song, Y. W. Choi, S. Y. Chi, K. K. Kim, and Y. K. Chung, "Scene text extraction in natural scene images using hierarchical feature combining and verification," *ICPR*, vol. 2, pp. 679–682, 2004.

[14] Wumo Pan, T. D. Bui, and C. Y. Suen, "Text detection from natural scene images using topographic maps and sparse representations," *IEEE ICIP*, 2009.