Text Detection in Urban Scenes

Sergio ESCALERA, Xavier BARÓ, Jordi VITRIÀ, and Petia RADEVA
Dept. Matemàtica Aplicada i Anàlisi, Gran Via 585, 08007, Barcelona
Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona
{sergio,xevi,jordi,petia}@maia.ub.es

Abstract. Text detection in urban scenes is a hard task due to the high variability of text appearance: different text fonts, changes in the point of view, or partial occlusion are just a few problems. Text detection can be specially suited for georeferencing business, navigation, tourist assistance, or to help visual impaired people. In this paper, we propose a general methodology to deal with the problem of text detection in outdoor scenes. The method is based on learning spatial information of gradient based features and Census Transform images using a cascade of classifiers. The method is applied in the context of Mobile Mapping systems, where a mobile vehicle captures urban image sequences. Moreover, a cover data set is presented and tested with the new methodology. The results show high accuracy when detecting multi-linear text regions with high variability of appearance, at same time that it preserves a low false alarm rate compared to classical approaches.

1. Introduction

Text recognition in outdoor scenes is one the most challenging problems in Computer Vision. On any artificial environment, as a city, a road, or a building, there is a large amount of textual information that we constantly use in order to navigate, locate a certain shop, or simply to decide which bottle of milk we buy. In general, the name of a product or a business is written using some stylistic font or combined with a representative image, allowing to be easily recognized by people. In this context, text regions adopt a wide variety of shapes, aspects, and sizes.

Efficient methods for text recognition in this scenario can be used in multiple applications, such as in mobile mapping systems to locate business in maps, self-positioning in navigation systems, tourists assistants [1], or systems to help visually impaired people to move in a city [2] and perform their daily tasks. Although most of recent works concern to text extraction for video indexing [3], the research on text extraction from natural scene images has been growing in the last years [4].

In the literature, there are many works related to locating text. First works were limited to the document analysis field, where text structure is often known, and the background is, in general, homogeneous. Lately, the increasing availability of devices with video recording capabilities and the increasing demand of video indexing have been reflected in a new generation of text detectors for video content.

There exist two main approaches related to the text detection problem:

• **Component-based:** In this case, text region is detected by analyzing the geometrical arrangement of detected components that belong to characters. Many methods to detect the components have been proposed, generally based on edges, greyscale/color homogeneity [5], or mathematical morphology operators [6]. Examples of this approach can be found in [7], where Smith detects text as horizontal rectangular structures of clustered sharp edges

• **Texture based**: Those approaches use texture to differentiate text region from background. Examples of this approach can be found in Jain works, where various textures are used to separate text, graphics, and halftone image regions in scanned grayscale document images [8]. Zhong utilizes the texture characteristics of text lines to extract text in grayscale images with complex backgrounds [8]. He locates candidate caption text regions directly in DCT compressed domain using the intensity variation information encoded in the DCT domain [8].

Each approach for text detection has different advantages/drawbacks concerning accuracy, efficiency, and computational requirements. For instance, component-based methods can locate text quickly but they have difficulties when the text is embedded in complex background or in contact with other graphical objects [9]. On the other hand, texture-based methods decrease the dependency on the text size, but they have difficulty to find accurate boundaries of text areas. All these works make evident that the text areas cannot be perfectly extracted from the image since natural scenes consist of complex objects, sometimes highly textured, such as buildings or trees, where it is common to obtain false text detections and misses.

In order to deal with the inherent previous problems of text detection in outdoor scenes, we present an approach for text detection in clutter scenes, where text appears with high variability of appearance. We use an object detection approach, based on a cascade of detectors, each one learnt using Gentle AdaBoost [10]. Each detector learns spatial information of simple gradient-based features and Census Transform images, which together encode the text structure. The system has been tested in the context of a mobile mapping application in order to locate interesting textual information in a map. Experiments either in highway sequences, urban sequences, as well as in a cover image data set have been performed, obtaining high robustness, speed, and a lower false detection rate compared to classical approaches.

The paper is organized as follows: In Section 2, gradient-based features, Census Transform images, and the learning procedure are described. Section 3 shows the experimental results in three new data sets: a highway data set, an urban data set, and a cover data set. Finally, Section 4 concludes the paper.

2. Gradient-based text features

The problem of automatic text detection in urban scenes is a hard task because of the high variability of text appearance (Fig. 1(a)). In order to deal with this problem, we need to look for discrimitative features able to distinguish between text and non-text. For this task, we propose to use a set of gradient-based and Census Transform features [11]. The images from which we compute the features are obtained from the text detection challenge of the ICDAR 2003 conference [12]. Some images from this data set are shown in Fig. 1(b).

(a) (b)

Figure 1. (a) Examples of text regions in urban scenes, and (b) some samples from the text data of [12].

Given a text region T, we compute its gradient vector as $\nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}\right)$, and its gradient magnitude as $|\nabla T| = \sqrt{\frac{\partial T}{\partial x}^2 - \frac{\partial T}{\partial y}^2}$. The Census Transform is obtained computing for each pixel a binary code of eight bits generated from neighboring intensity comparisons. This binary code is then converted to a 10-base number in [0,..,255]. In Figure 2 an example of a census transform value CT for a given pixel is shown. If the intensity of the analyzed pixel is greater or equal than its neighbor, the position from the binary code is set to one. Then, the output bits are collected from top to bottom and from left to right. Finally, a CT value of 214 is obtained for the analyzed pixel of the example. In Fig. 3, the T, $\frac{\partial T}{\partial x}$, $\frac{\partial T}{\partial y}$, $|\nabla T|$, and CT images for an input text region are shown.

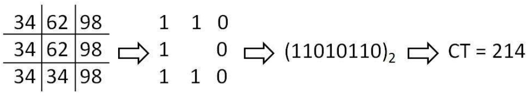

Figure 2. Example of a CT pixel value computation.

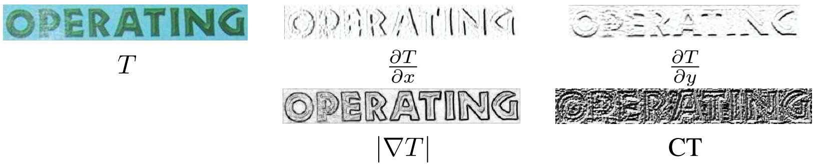
T
$\frac{\partial T}{\partial x}$
$\frac{\partial T}{\partial y}$
$|\nabla T|$
CT

Figure 3. Gradient-based and CT text regions.

In order to determine if these types of regions generalize the text structure, we compute the mean of the four terms $|\frac{\partial T}{\partial x}|$, $|\frac{\partial T}{\partial y}|$, $|\nabla T|$, and CT considering the magnitude for all the selected text samples from the data set (we select about 1000 text regions). The obtained results are shown in Fig. 4. One can see a general visual discriminative structure for each of the four operators. In particular, the CT operator shows a structure similar to that one obtained by the gradient operator. The use of oriented-based features for text detection was originally proposed by [13], where text is described based on a reduced set of $\frac{\partial T}{\partial x}$ and $\frac{\partial T}{\partial y}$ components. In this paper, we analyze its extension including the CT operator.

Defining a text region R of height h and width $w = 2 \cdot h$, we consider the five sub-partitions $\{R_1, .., R_5\}$ shown in Fig. 5, being r_i^j the ith sub-region of the jth sub-partition of R. Then, we define the set of features of Table 1, where $r_{i,k}^j$ is the value of the kth pixel in the ith sub-region of the jth sub-partition of R. The last column of the table

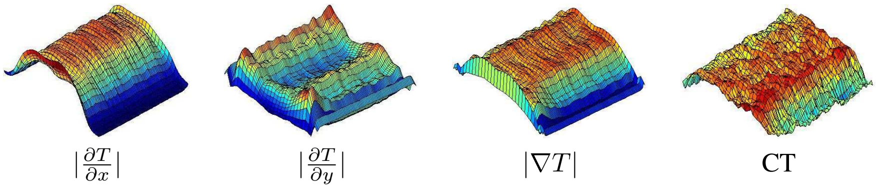
$|\frac{\partial T}{\partial x}|$
$|\frac{\partial T}{\partial y}|$
$|\nabla T|$
CT

Figure 4. Mean values of the four operators over text regions.

R_1	$\sum_k r^1_{2,k} - \sum_k r^1_{1,k}$ $\sum_k r^1_{2,k} - \sum_k r^1_{3,k}$ $\sum_k r^1_{2,k} - (\sum_k r^1_{1,k} + \sum_k r^1_{3,k})$	$r^1_1 : 1/6h \times 2h$ $r^1_2 : 2/3h \times 2h$ $r^1_3 : 1/6h \times 2h$
R_2	$\sum_k r^2_{2,k} - (\sum_k r^2_{1,k} + \sum_k r^2_{3,k})$ $\sum_k r^2_{4,k} - (\sum_k r^2_{3,k} + \sum_k r^2_{5,k})$ $\sum_k r^2_{2,k} + \sum_k r^2_{4,k} - (\sum_k r^2_{1,k} + \sum_k r^2_{3,k} + \sum_k r^2_{5,k})$	$r^2_1 : 1/9h \times 2h$ $r^2_2 : 1/3h \times 2h$ $r^2_3 : 1/9h \times 2h$ $r^2_4 : 1/3h \times 2h$ $r^2_5 : 1/9h \times 2h$
R_3	$\sum_k r^3_{2,k} - \sum_k r^3_{1,k}$ $\sum_k r^3_{2,k} - \sum_k r^3_{3,k}$ $\sum_k r^3_{2,k} - (\sum_k r^3_{1,k} + \sum_k r^3_{3,k})$	$r^3_1 : h \times 2/3h$ $r^3_2 : h \times 2/3h$ $r^3_3 : h \times 2/3h$
R_4	$\sum_k r^4_{2,k} - \sum_k r^4_{1,k}$	$r^4_1 : h/2 \times 2h$ $r^4_2 : h/2 \times 2h$
R_5	$\sum_k r^5_{2,k} - \sum_k r^5_{1,k}$	$r^5_1 : h \times h$ $r^5_2 : h \times h$

Table 1. Gradient-based features.

stands for the size of the sub-regions. These features are computed over the four images $|\frac{\partial T}{\partial x}|$, $|\frac{\partial T}{\partial y}|$, $|\nabla T|$, and CT for each text region R, representing a total of 44 gradient and CT features per text region.

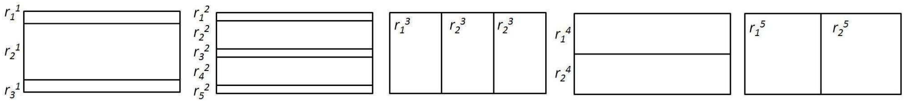

Figure 5. Sub-partitions $\{R_1, .., R_5\}$ considered for a text region R.

In order to learn the previous set of features, we train a cascade of classifiers [10]. In particular, we use the Gentle version of Adaboost as the classifier, and Decision Stump as the Adaboost weaklearner. In order to speed up the detection procedure, the classifiers are included in a cascade [10], which learns the feature space using the integral image representation [10]. At the detection step, the cascade of classifiers is tested over a set of rectangles at different sizes be means of a windows slicing procedure with a windows size of $w = 2 \cdot h$. Since a text region will be composed by a set of positive region detections, we apply a post-processing step. In order to define the text region and discard false positives, we count the number of regions that falls in the connected area defined by the union of all detected regions. If this number of regions is higher than a given threshold parameter T_h, then, the bounding box that contains the connected detected regions defines the final multi-linear text region. With this procedure, rotated text or affine distortions of the text because of changes in the camera point of view can also be detected.

3. Results

In order to present the results, first we discuss the data, methods, validation, and experiments.

• *Data*: We used the video sequences obtained from the Mobile Mapping System of [14] to test the text detection methodology on outdoor scenes. The video sequences correspond to highway and urban scenes from Barcelona. In this system, the position and orientation of the different traffic signs are measured with video cameras fixed on a moving vehicle 6. We also designed a cover data set composed by 6000 cover images. For each image, a xml file contains information about the text bounding boxes and the url where the image was found. Some example of this data set are shown in Figure 7.[1]

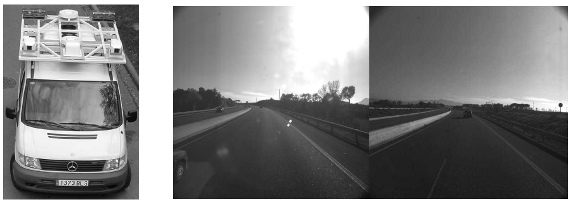

Figure 6. Mobile mapping vehicle and examples of captured images at different conditions.

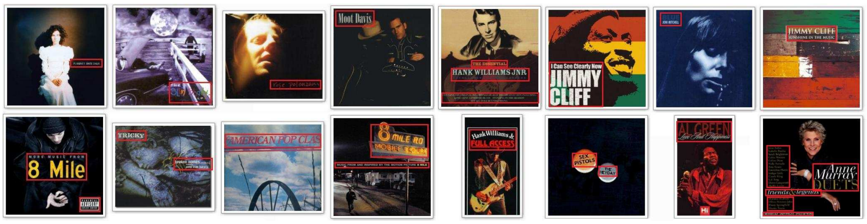

Figure 7. Cover data set samples with labeled text regions.

• *Methods*: We train a Cascade of classifiers using 200 runs of Gentle Adaboost with decision stumps. 1000 positive text regions from the text detection challenge of the ICDAR 2003 conference [12] were used to learn ten levels of the cascade, taking 2000 negative samples per level from the negative set of Mobile Mapping road frames and Google random background images. The performance settings were a minimum accuracy of 99.8% and a maximum false alarm of 40% per level. The cascade achieves a final theoretical hit ratio of $\simeq 0.98$ and a theoretical false alarm rate near $\simeq 4 \cdot 10^{-5}$ over the training data. Two cascades were trained with the previous settings for two different feature sets. The first one considers just the gradient-based features as defined in [13] (33 features), and the second one uses the gradient and CT features described in this paper (44 features).

• *Validation*: We define a minimum region of 28×56 pixels resolution to apply the cascade, using an initial horizontal and vertical displacement of five pixels, and increasing the scale by factor 1.2. The resolution of the urban images is of size 400×600. The

[1] These data sets and ground truths are publicly available under request to the authors of this paper.

accuracy A is determined as $A = \frac{\text{\# Hits}}{\text{\# Text regions in the analyzed frames}}$, where the hits correspond to the number of detected text regions, and the hits are obtained if the intersection between the ideal region and the detected region is at least a 60% of the size of the highest region, similar to the validation procedure of [15]. The false alarm rate FR is determined as $FR = \min\left(\frac{\text{\# False positive detections}}{\text{\# Analyzed frames}}, 1\right)$. In the case of the evaluation of the cover data set we use the metric evaluation used in [16]:

$$Performance = \frac{\text{Detected text area} \cap \text{ground truth text area}}{\text{Detected text area} \cup \text{ground truth text area}} \tag{1}$$

• *Experiments*: We split the experiments in three types: First, detecting text from information panels in highway sequences, second, detecting arbitrary text from urban scenes, and third, validating our approach over the presented cover image data set.

3.1. Text in highway sequences

For this experiment, we tested the trained cascades in a video sequence of 2000 frames, corresponding to $\simeq$ 20km of road, from which 217 frames contain text regions. We consider the appearance of a text region at each frame as an independent text region, though it may correspond to the same information panel. In this sequence, the text basically appears in frontal view information panels, with few distortions and similar fonts. The results obtained in this experiment are an accuracy of $A = 0.89$ and a false alarm rate of $FR = 0.02$, compared to an accuracy of $A = 0.79$ and a false alarm rate of $FR = 0.17$ obtained by the cascade of gradient-based features. In our case, an accuracy near 90% is considerably high given the difficulty of the problem and the results reported in literature (i.e. the participants of the ICDAR 2003 challenge reported at most results upon 60% for similar images to ours). A false alarm rate of 2% is considerably low. Notice that only in a 2% of the frames a positive region appears and that for each frame of 400×600 pixels, we analyze the following number of windows: $W = \sum_{s \in [1,1.2,..,5.15]} \frac{1}{s}\left(\frac{400-28 \cdot s}{5} \cdot \frac{600-56 \cdot s}{5}\right)$, which corresponds to $\simeq 35000$ analyzed regions per frame. Moreover, the C++ implementation of the procedure spends $\simeq 1$ second per frame. Some detection results are shown in Fig. 8. The marked regions correspond to isolate region detections. The final detections are shown in the top of each image. In some cases, the advertisement text of the cars is detected (if it is at least of the minimum size of 28×56 pixels). Notice that as we consider a minimum of ten connected isolated detections to define a multi-linear text region detection, short isolated words are not detected, since it would also considerable increase the false alarm rate. Finally, the last image of Fig. 8 shows a false positive detection. One can see on the top of this image that the content of this false positive region has a visual structure close similar to a possible text instance.

3.2. Text in urban scenes

For this experiment, we tested the trained cascades in a video sequence of 2000 frames, corresponding to $\simeq$ 8km of road, from which 487 frames contain text regions. We consider the appearance of a text region at each frame as an independent text region, though it may correspond to the same real text region. In this sequence, the text appears with

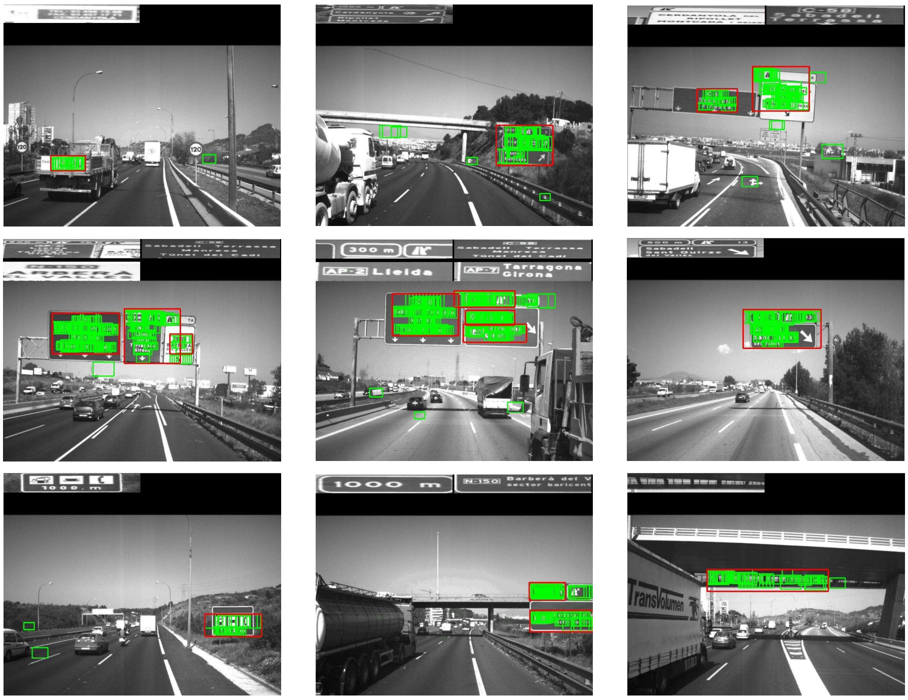

Figure 8. Positive text detections from highway images. In this case the text appears in similar conditions. Some text regions corresponds to cars' advertising, and most of the text regions are from information panels. In the top of each image, the content of the detected text regions is shown. The last image shows a false positive detection where the inner structure of the bounding box is close similar to a possible text instance.

affine distortions, different font types, or even occluded, being a non-controlled text detection problem in a non-controlled environment. The results obtained in this experiment are an accuracy of $A = 0.78$ and a false alarm rate of $FR = 0.05$, compared to an accuracy of $A = 0.63$ and a false alarm rate of $FR = 0.29$ obtained by the cascade of gradient-based features. In our case, a detection rate $\simeq 80\%$ is considerably high taking into account the high variability of text appearance in urban scenes. Examples of detections are shown in Fig. 9. One can see that by determining the multi-linear text region by a set of connected detections, we are able to find text that suffers from rotation or irregular deformations. The last image in Fig. 9 shows a false positive detection, which visually could be confused with a string (i.e. 'HHHHHHHHHHH').

The experimental values obtained for A and FR varying the number of considered connected regions T to predict a positive text region are shown in Fig. 10 for the cascade considering gradient and CT based features. Notice that when reducing the value of T more positive regions are detected. However, the number of false positive detections dramatically increases. The number of $T = 10$ used for these experiments was experimentally selected as a trade-off between both A and FR values.

Finally, Fig. 11 shows an analyzed urban sequence using the cascade of gradient and CT based features. This sequence shows detections from near consecutive frames. One can see that the text from shops is tracked among frames and updated with the new width and height. Moreover, text is detected with occlusions, high rotations, and different font

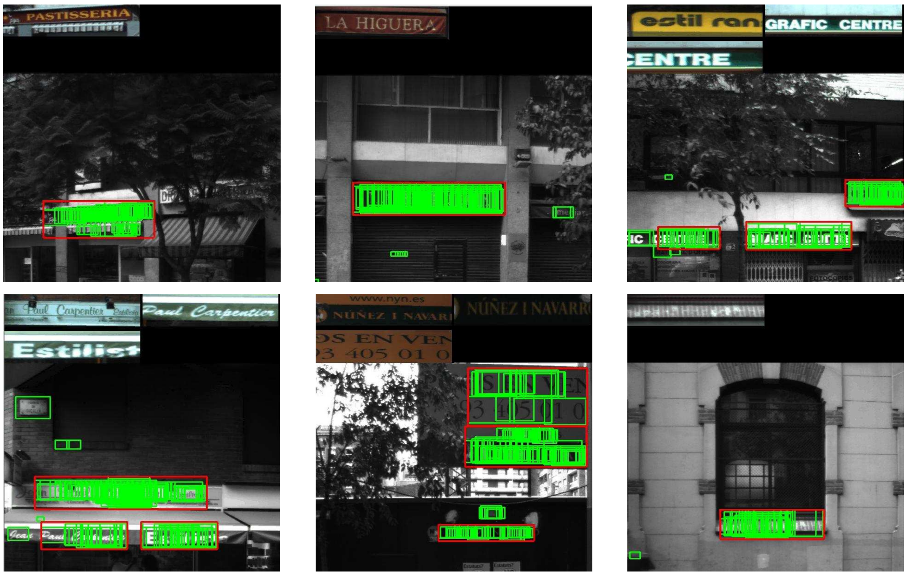

Figure 9. Samples of text detection regions from urban scenes. The detections consider high variability of text appearance. The last detection shows a false text region detection, where the detected region can be confused with a string (i.e. 'HHHHHHHHH'). In the top of each image, the content of the detected text regions is shown.

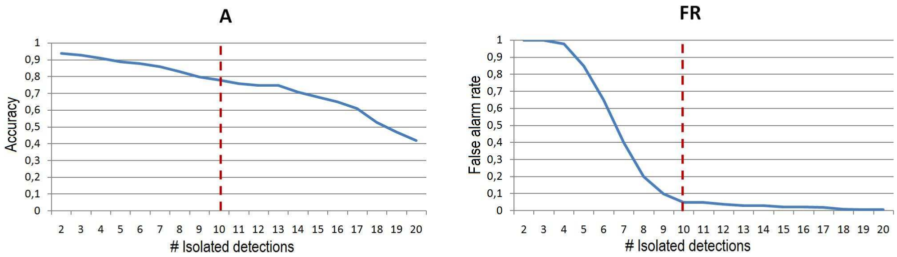

Figure 10. Accuracy and false alarm rate for different values of connected components.

types. Notice that the final detections are only those marked with a red bounding box (as shown in the top of the images) and not the green regions which correspond to isolated positive regions, though most of them contain small text or short words.

3.3. Cover data set text detection

For this experiment, we tested the trained cascades over the 6000 Cover data set samples. The results using the performance metric of eq.(1) are a performance of 0.63 compared to a performance of 0.45 obtained by the cascade of gradient-based features. In our case, an accuracy upon 60% is considerably high given the pessimistic metric of eq.(1).

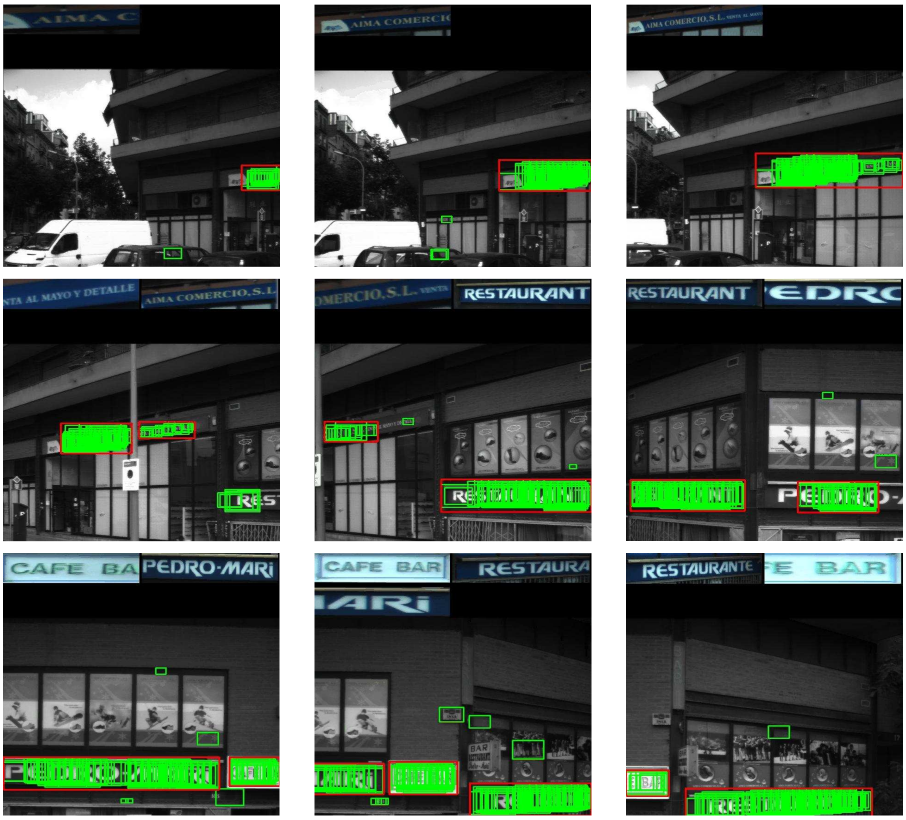

Figure 11. Text detection in a video sequence of urban frames.

4. Conclusions

We presented a fast and robust methodology to deal with the problem of text detection in urban scenes. Text regions are globally described using a set of gradient and Census Transform based features. Then, a cascade of classifiers detects text regions. With this approach, text with high variability of appearance because of changes in the point of view, rotation, occlusions, or different font types is detected. The methodology is applied over the sequences obtained by a Mobile Mapping system. Moreover, we presented a cover image data set where we also validated our methodology. The results show high accuracy detecting multi-linear text regions with high variability of appearance, at same time that it preserves a lower false alarm rate than classical approaches.

5. Acknowledgements

This work has been performed in collaboration with the Institut Cartogràfic de Catalunya. This work has been partially supported by the projects TIN2006-15694-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

[1] J. Yang, J. Gao, Y. Zang, X. Chen, A. Waibel, An automatic sign recognition and translation system, Workshop on Perceptive User Interfaces.

[2] N. Ezaki, M. Bulacu, L. Schomaker, Text detection from natural scene images: towards a system for visually impaired persons, Pattern Recognition (2004) 325–330.

[3] H. Yan, Y. Zhang, Z. Hou, M. Tan, Automatic text detection in video frames based on bootstrap artificial neural network and ced, Central Europe on Comp. Graphics, Visualization and Computer Vision.

[4] D. Doermann, J. Liang, H. Li, Progress in camerabased document image analysis, Document Analysis and Recognition (ICDAR 2003) 1 (2003) 606–616.

[5] Y. Zhong, K. Karu, A.K.Jain, Locating text in complex color images, PR 28 (1995) 1523–1536.

[6] L. Gu, N. Tanaka, T. Kaneko, R. Haralick, The extraction of characters from cover images using mathematical morphology, Tran. of The Inst. of Electronics, Inf. and Communication Engineers of Japan.

[7] M. Smith, T. Kanade, Video skimming and characterization through language and image understanding techniques, technical report.

[8] Z. Yu, Z. Hongjiang, A. Jain, Automatic caption localization in compressed video, IEEE Trans. On PAMI 22 (2000) 385–392.

[9] J. Xi, X.-S. Hua, X.-R. Chen, L. Wenyin, H.-J. Zhang, A video text detection and recognition system, IEEE International Conference on Multimedia and Expo.

[10] P. Viola, M. Jones, Robust real-time object detection, IJCV.

[11] R. Rabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, ECCV 2 (1994) 151–158.

[12] I. . conference data.
URL http://algoval.essex.ac.uk/icdar/Competitions.html

[13] C. Xiangrong, A. Yuille, Detecting and reading text in natural scenes, Computer Vision and Pattern Recognition 2 (2004) 366–373.

[14] R. Alamús, A. Baron, E. Bosch, J. Casacuberta, J. Miranda, M. Pla, S. Sànchez, A. Serra, J. Talaya, On the accuray and performance of the geomobil system, ISPRS.

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. V. Gool, A comparison of affine region detectors, IJCV 65 (1/2) (2005) 43–72.

[16] T. Retornaz, B. Marcotegui, Scene-text localization based on ultimate opening, International Symposium on Mathematical Morphology.