Image and Vision Computing 28 (2010) 1413–1426

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A two-stage scheme for text detection in video images

Marios Anthimopoulos *, Basilis Gatos, Ioannis Pratikakis

Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", 153 10 Athens, Greece

ARTICLE INFO

Article history:
Received 16 July 2009
Received in revised form 17 February 2010
Accepted 3 March 2010

Keywords:
Text detection
Video OCR
Content-based indexing
SVM

ABSTRACT

This paper proposes a two-stage system for text detection in video images. In the first stage, text lines are detected based on the edge map of the image leading in a high recall rate with low computational time expenses. In the second stage, the result is refined using a sliding window and an SVM classifier trained on features obtained by a new Local Binary Pattern-based operator (eLBP) that describes the local edge distribution. The whole algorithm is used in a multiresolution fashion enabling detection of characters for a broad size range. Experimental results, based on a new evaluation methodology, show the promising overall performance of the system on a challenging corpus, and prove the superior discriminating ability of the proposed feature set against the best features reported in the literature.

1. Introduction

The tremendous increase of multimedia content has raised the need for automatic semantic information indexing and retrieval systems. Many methods have been proposed for the semantics extraction of various granularity levels from audiovisual content. Textual information in videos proves to be an important source of high-level semantics. There exist mainly two kinds of text occurrences in videos, namely artificial and scene text. Artificial text, as the name implies, is artificially added in order to describe the content of the video or give additional information related to it. This makes it highly useful for building keyword indexes. Scene text is textual content that was captured by a camera as part of a scene such as text on T-shirts or road signs and usually brings less related to video information. In Fig. 1, yellow boxes denote artificial text while red boxes delimit the scene text. Text can also be classified into normal or inverse. Normal is denoted any text whose characters have lower intensity values than the background while inverse text is the opposite [1]. In Fig. 2, "EURO" is inverse while "SPORT" is normal text. The procedure of textual information extraction from videos is usually split into four distinct steps: (i) detection, (ii) tracking, (iii) segmentation and (iv) recognition. Among all steps, detection step is the most crucial and although it has been extensively studied in the past decade presenting quite promising results, there are still challenges to meet. Further down we will discuss the different approaches used towards the text detection problem, the corresponding drawbacks and the remaining challenges. In this discussion we will focus on the methods designed for detecting artificial text in video images. For scene text detection in camera-based images the reader should refer to the survey papers [2–4].

Most of proposed text detection methods use as representative text features, color, edge and texture information. To exploit this information, i.e. describe text and discriminate it from the background, some researchers apply heuristic rules derived by empirical constraints while others use machine-learning methods trained on real data. Recently, some hybrid approaches have been proposed.

Many existing heuristic methods, derived from document analysis research area, are based on color or intensity homogeneity of characters. They detect character regions in the image and then group them into words and text lines based on geometrical constraints. These methods, also known as connected component (CC) methods, can perform satisfactorily only on high quality images with simple background and known text color, assumptions that usually do not apply in the case of video images. Moreover, text in video images often suffers from color bleeding due to video compression. Typical CC approaches can be found in [5,6].

Some other heuristic methods detect text based on edge information, i.e. strength, density or distribution. Sato et al. [7] apply a 3×3 horizontal differential filter to the entire image with appropriate binary thresholding followed by size, fill factor and horizontal–vertical aspect ratio constraints. Xi et al. [8] propose an edge-based method based on an edge map created by Sobel operator followed by smoothing filters, morphological operations and geometrical constraints. Cai et al. [9] and Lyu et al. [10] suggest the use of local thresholding on a sobel-based edge strength map. Anthimopoulos et al. [11] use Canny edge map followed by morphological operations and projection analysis. Kim and Kim [12] instead of using an explicit

* Corresponding author. Tel.: +30 210 650 3218; fax: +30 210 653 2175.
E-mail addresses: anthimop@iit.demokritos.gr (M. Anthimopoulos), bgat@iit.demokritos.gr (B. Gatos), ipratika@iit.demokritos.gr (I. Pratikakis).

0262-8856/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2010.03.004

Fig. 1. Example of artificial and scene text. Yellow boxes bound artificial text while red indicate scene text.

edge map as an indicator of overlay text, they suggest the use of a transition map generated by the change of intensity and a modified saturation. A heuristic rule based on the different Local Binary Patterns (LBP) is used for verification. These heuristic techniques proved to be very efficient and satisfactory robust for specific applications with high contrast characters and relatively smooth background. However, the fact that many parameters have to be estimated experimentally condemns them to data dependency and lack of generality.

DCT coefficients of intensity images have been widely used as texture features and have also been used for heuristic text detection methods [13–16]. The DCT coefficients globally map the periodicity of an image and can be a quite efficient solution for jpeg and mpeg encoded images and videos. In this case, the pre-computed coefficients of 8×8 pixel block units are used. However, this block size is not a large enough area to sufficiently depict the periodical features of a text line and the computation of DCT for larger windows even by the fast DCT transform proves quite costly. In addition, these methods still use empirical thresholds on specific DCT-based features and therefore they lack adaptability.

Several machine learning-based approaches have been proposed for the detection of text areas with great success. These approaches are based on sliding windows that scan the image and machine learning techniques which classify each window as text or non-text. Machine learning classifiers have proved to be an appealing solution for many problems that cannot be defined in a strict mathematical manner. Jung [17] and Kim et al. [18] directly use the color and gray values of the pixels as input for a neural network and an SVM, respectively. Wolf and Jolion [19] use an SVM trained on differential and geometrical features. Lienhart and Wernicke in [20] used as features the complex values of the gradient of the RGB input image fed to a complex-valued neural network. Li et al. [21] suggest the use of the mean, second order (variance) and third-order central moments of the LH, HL, and HH component of the first three levels of each window to train a three-layer neural network. Zhang et al. [22] proposed a system for object detection based on Local Binary Patterns (LBP) and Cascade histogram matching. They applied the proposed method to video text and car detection. The main shortcoming of the methods attributed to this category is the high computational complexity since a sliding window is required to scan the entire image with a typical step of 3 or 4 pixels, demanding thousands of calls to the classifier per image.

Fig. 2. Example of inverse and normal text.

Recently, some hybrid methods have also been proposed, that combine the efficiency of heuristic methods with machine-learning accuracy and generalization. These methods usually consist of two stages. The first localizes text with a fast heuristic technique while the second verifies the previous results eliminating some detected area as false alarms using machine learning. In [23], Chen et al. use a localization/verification scheme which claim to be highly efficient and effective. For the verification part, Constant Gradient Variance (CGV) features are fed to an SVM classifier. Ye et al. [24] propose a coarse-to-fine algorithm based on wavelets. The first stage applies thresholding on the wavelet energy in order to coarsely detect text, while the second identifies the coarse results using an SVM and a more sophisticated wavelet-based feature set. Jung et al. [25] apply as a first stage, a stroke filtering and they also verify the result using an SVM with normalized gray intensity and CGV features. Then, a text line refinement module follows, consisting of text boundary shrinking, combination and extension functions. However, the machine learning verification task used by these methods can only take a binary decision, i.e. if an initial result is text or not without having the capability to refine it. For example, if the resulting bounding box of the first stage contains text as well as background pixels, in the second stage it will be either entirely verified as text, or discarded as false alarm.

Concluding the discussion of the previous works, we can say that although the specific research area has shown a great progress, there are still challenges to be considered. Machine-learning methods have shown important capabilities for generalization but proved to be computationally expensive. Hybrid methods benefit regarding efficiency but they are actually based on the initial results of heuristic methods which fail to deal with complex background. Another great challenge is the choice of the feature set used by machine learning techniques to discriminate text from non-text areas. The features used until now, usually inherited from texture segmentation research area, are not capable to adapt to the specific problem and fail to discriminate varying-contrast text from high-contrast background. Finally, another essential factor for the optimization of every text detection algorithm is the choice of an objective evaluation methodology.

In this work we propose a two-stage approach with a novel machine learning refinement which cannot only verify the initial result but refine its boundaries as well. In that way the whole system's performance is based on this refinement, instead of relying to the initial heuristic results. This machine learning stage uses a new, highly discriminative feature set, derived from a proposed operator that captures the edge structure for different levels of contrast, and has shown superior performance against other features reported in the literature. Finally, a novel evaluation methodology is introduced which is based on the estimated number of