Expert Systems with Applications 42 (2015) 5554–5566

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Bayesian classifier for multi-oriented video text recognition system

Sangheeta Roy [a], Palaiahnakote Shivakumara [a,*], Partha Pratim Roy [b], Umapada Pal [c], Chew Lim Tan [d], Tong Lu [e]

[a] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lampur, Malaysia
[b] Indian Institute of Technology, Roorkee, India
[c] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
[d] School of Computing, National University of Singapore, Singapore
[e] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

ARTICLE INFO

Article history:
Available online 14 March 2015

Keywords:
Wavelet and gradient sub-bands
Bayesian classifier
Video text lines
Video text binarization
Video text recognition

ABSTRACT

Developing an automatic system for recognizing video texts such as signboards, street names, room numbers, building names and hotels names is challenging due to low resolution, complex background, font or font size variations, and multiple orientations of texts. In this paper, we develop a new system to recognize video texts through binarization by introducing a Bayesian classifier. We explore wavelet decomposition and gradient sub-bands to enhance text information in video. The enhanced information is used in different ways to calculate the requirement of Bayesian classifier, such as a priori probability and conditional probabilities of text pixels to estimate the posterior probability automatically, which results in text components. Connected component analysis is then applied to restore missing text information before sending it to an OCR engine if any disconnection exists in the text components. Experimental results on video data, the benchmark ICDAR scene character data (camera images) and arbitrary orientation data (camera images) show that the proposed method outperforms existing baseline methods in terms of recognition rates at both character and pixel levels.

1. Introduction

With the recent progresses in computer science and technology, especially in the evolutions in TV, internet and mobile systems, the use of video or scene images for daily activities of human beings increases drastically. As a result, there is a demand for developing systems which can understand high level semantics of both video and images to cope with the requirements in daily activities. However, conventional systems like content based image retrieval systems may not give satisfactory results due to lack of semantics to bridge the gap between low level and high level features (Doermann, Liang, & Li, 2003; Lyu, Song, & Cai, 2005; Zang & Kasturi, 2008). To overcome this problem, researchers have made attempts to develop systems for recognizing text in video to understand the content at high level, which can help us to retrieve video events according to semantics. Besides, the involvement of text recognition plays a vital role in developing various real time applications, such as assisting blind people to walk freely on streets without aid, assisting tourists to reach their destinations, assisting safe driving and retrieving exciting sports events from large video databases (Crandall, Antani, & Kasturi, 2003; Doermann et al., 2003; Lyu & Song, 2005; Zang & Kasturi, 2008) efficiently and accurately. In addition, a system that has the ability to read characters, signs and messages would undoubtedly be a very useful complement for navigation, tracking and surveillance applications like tracking text in movies, locating cars based on recognizing license plate numbers etc. It is noted from the papers (Grafmuller & Beyerer, 2013; Park & Kim, 2013; Risnumawan, Shivakumara, Chan, & Tan, 2014; Wei & Lin, 2012) that developing such an expert recognition system which meets the requirement of real time industry is really challenging and hard for researchers due to complex background, font or font size variations, rotations, etc.

It is known that video consists of two types of texts: graphics text and scene text. Graphics text is manually added, and scene text exists naturally as a part of an image. Since graphics text is edited, it has good contrast and clarity to make it readable, while scene text is a part of the image and generally possesses undesirable characteristics which hamper its recognition. Therefore, there is a need for developing a system which can recognize text with good accuracies for both graphics and scene texts in video

* Corresponding author.
E-mail addresses: 2sangheetaroy@gmail.com (S. Roy), shiva@um.edu.my (P. Shivakumara), 2partharoy@gmail.com (P.P. Roy), umapada@isical.ac.in (U. Pal), tancl@comp.nus.edu.sg (C.L. Tan), lutong@nju.edu.cn (T. Lu).

http://dx.doi.org/10.1016/j.eswa.2015.02.030
0957-4174/

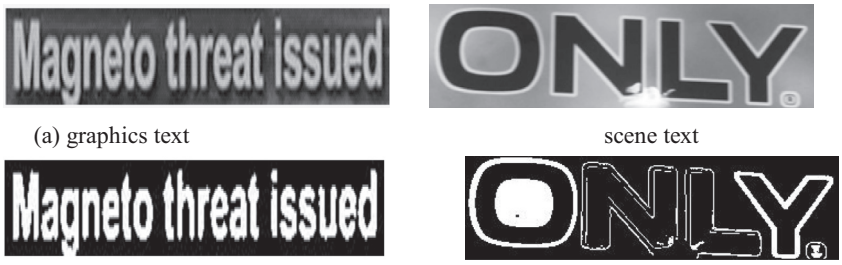

Fig. 1. Illustration for video text recognition through binarization.

(Crandall et al., 2003; Doermann et al., 2003; Lyu & Song, 2005; Zang & Kasturi, 2008).

It is true that the problem of text recognition is not new for the document analysis community as we can see many sophisticated methods for recognizing text with more than 95% recognition rate are reported in the literature. There are Optical Character Recognizers (OCR) available publicly for recognizing texts in documents and hence it is considered as a successful application in the field of pattern recognition and artificial intelligence (Aradhya, Hemantha Kumar, & Noushat, 2008; Jung, Kim, & Jain, 2004; Niblack, 1986; Otsu, 1979; Sharma, Pal, & Blumenstein, 2012; Wolf, Michel, & Chassaing, 2002). The features proposed in the OCR methods give good recognition rates when the shape of a character is well preserved. Since the text in scanned documents or camera based documents has a high resolution and a high contrast, text shape can be preserved without much difficulty. However, since video suffers from complex background and low resolution, it is hard to preserve the shape of every character. Therefore, the existing document analysis based methods may not be suitable for video text recognition.

In order to overcome the above problems, there are several methods proposed in the literature, such as the methods for recognizing text in natural scene images captured by a high resolution camera (Phan, Shivakumara, Lu, & Tan, 2013; Phan, Shivakumara, Tian, & Tan, 2013; Shi, Wang, Xiao, Gao, & Hu, 2014; Yao, Bai, & Liu, 2014; Ye & Doermann, 2014; Yi & Tian, 2014). According to the literature, we can find two ways to solve this problem, which are: (1) recognizing text by proposing their own features and classifiers, which generally do not use binarization and the available OCR. In other words, they develop a separate OCR for recognizing video text. (2) Recognizing text by developing a robust binarization method such that the available OCR can be used for recognition in video. The former one is too expensive and has its own limitations, such as the use of a classifier and the training samples, which restrict the ability to adopt it for different scripts, data and applications. On the other hand, the latter one is inexpensive compared to the former one as it makes use of the available OCR. Therefore, we prefer the latter one to solve the video text recognition problem in this work rather than developing a separate OCR. Further, since text in image has high contrast, the existing methods take this advantage to extract features directly or indirectly based on the shapes of characters. As a result, though these methods solve the complex background problem, they still are not smart enough to solve the video text recognition problem due to the presence of both graphics and scene texts, which additionally suffer from both complex background and low resolution.

It is evident from natural scene character recognition methods (Chen & Odobez, 2005; Lyu & Song, 2005; Neumann & Matas, 2010) that a document OCR engine does not work for camera based natural scene images due to the failure of binarization in handling non-uniform background and non-uniform illumination. This shows that despite high contrast of camera images, so far, the best accuracy reported is 67% for ICDAR-2003 competition data (Neumann & Matas, 2010). It is also noted that character recognition rate varies from 0% to 45% (Chen & Odobez, 2005) if we apply OCR directly on video text, which is much lower than scene character recognition accuracy. Hence, there is a need for developing a powerful recognition system which can give good recognition rate for video text. One example to illustrate the problem of video text is shown in Fig. 1, where the existing binarization method (Roy, Shivakumara, Roy, & Tan, 2012) recognizes graphics text correctly, but the same method fails to recognize scene text correctly because the method does not preserve character shape for the scene text. This shows that conventional binarization methods work well for graphics text of high contrast with plain background, but not for scene text because it suffer from the effects of illumination variations that are unpredictable.

2. Related work

It is noticed from the above discussions on text recognition in natural scene images that there are two ways to achieve a good recognition rate for natural scene text. In the same way, according to the literature on video text recognition, we can find three ways for solving text recognition in video: (1) recognition using temporal information, which generally focuses on how to utilize temporal frames for achieving a good recognition rate since video provides temporal frames. For instance, Chen and Odobez proposed video text recognition using sequential Monte Carlo and error voting methods (Chen & Odobez, 2005). This method approximates the posteriori distribution of segmentation thresholds of text pixels in an image by a set of weighted samples. The set of samples is initialized by applying a classical segmentation algorithm on the first video frame and further refined by random sampling under a temporal Bayesian framework. This method is sensitive to thresholds when complex background is present in the image. Chen et al. proposed text detection and recognition in images and video frames based on a multiple hypothesis framework (Chen, Odobez, & Bourlard, 2004). The method considers the outputs of text detection step for recognition. Therefore, the performance of the method depends on the text detection step. In addition, determining multiple hypotheses for different situations is not as easy as defining the rules for scanned document text.