Pattern Recognition 52 (2016) 433–447

Contents lists available at ScienceDirect

Pattern Recognition

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

Fractional poisson enhancement model for text detection and recognition in video frames

Sangheeta Roy [a], Palaiahnakote Shivakumara [a], Hamid A. Jalab [a], Rabha W. Ibrahim [b], Umapada Pal [c], Tong Lu [d,*]

[a] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[b] Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur, Malaysia
[c] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
[d] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

ARTICLE INFO

Article history:
Received 11 April 2015
Received in revised form
13 October 2015
Accepted 15 October 2015
Available online 23 October 2015

Keywords:
Text detection
Text recognition
Laplacian operation
Fractional Poission model
Text enhancement

ABSTRACT

Performing Laplacian operation on video images is a common technique to improve image contrast to achieve good text detection and recognition accuracies. However, it is a fact that when Laplacian operation enhances contrast, at the same time it introduces too many noises. To alleviate this, the existing methods propose different enhancement methods and filters. In this paper, we propose a generalized enhancement model based on fractional calculus to increase the quality of images obtained by Laplacian operation. The proposed method considers edges and their neighbor information to derive a mathematical model for enhancing low contrast information in video as well as in scene images. Experimental results of text detection and recognition methods on different databases show that the proposed enhancement model improves their accuracies significantly. The enhancement model is compared with standard enhancement models to show that the proposed model outperforms the existing models in terms of quality measures. The usefulness of the proposed model is validated through text detection and recognition experiments.

1. Introduction

Day by day video text detection and recognition is receiving greater attentions by researchers with the aim of improving the performances of the existing text detection and recognition methods because real time applications like the systems for assisting a blind person to walk freely on roads, safely driving, and tracking license plates of moving vehicles, often require more than 90% detection and recognition accuracies [1–5] for security and surveillance purposes. However, achieving such a high accuracy is an elusive goal for researchers because video images suffer from degradations severely, which are caused by motion blur, lighting, non-uniform illumination, text movements and complex background [3,6,7]. To overcome these problems, existing methods [8–13] have been proposed in literature for enhancing text information in video images based on gradient operation with Laplacian mask because Laplacian helps in identifying abrupt changes from background to foreground and vice versa providing high positive and negative peaks. This is useful for both text detection and recognition as this information is the basis for extracting features to detect text and separate foreground (text) from background in binarization. For example, Shivakumara et al. [10] used high positive and negative peaks for segmenting the words in each text line in video. Phan et al. [11] used the transition from background to foreground and vice versa for text candidate selection to detect text in video images.

Similarly, the Laplacian operation has been used for binarization and recognition of text in video or images [12,13]. It is true that Laplacian helps in enhancing text information and text separation; however, it introduces too many noises while performing Laplacian operation over the image. To get rid of this problem, the existing methods [14–17] usually propose different criterion based on filters to remove noise effect caused by Laplacian operation. It is evident from the following methods. Shivakumara et al. [14] proposed wavelet and color features for text detection in video, where an enhanced image obtained by the combination of R, G and B color spaces is considered as the input for Laplacian operation to avoid noise effect in the selection of text candidates. Shivakumara et al. [16] also proposed a Laplacian approach for multi-oriented text detection in video, where Fourier

* Corresponding author.
E-mail addresses: 2sangheetaroy@gmail.com (S. Roy), shiva@um.edu.my (P. Shivakumara), hamidjalab@um.edu.my (H.A. Jalab), rabhaibrahim@um.edu.my (R.W. Ibrahim), umapada@isical.ac.in (U. Pal), lutong@nju.edu.cn (T. Lu).

http://dx.doi.org/10.1016/j.patcog.2015.10.011
0031-3203/

Fig. 1. Text detection and recognition results for an input video image, its Laplacian image and enhanced image. (a) Text detection by the text detection method [14] for the input, the Laplacian and the enhanced images. (b) Text line images chosen from respective results in (a). (c) Binarization results of the method [18] for respective text line images in (b). (d) Recognition results of the OCR engine [19] for the respective results in (c).

transform is proposed as an ideal low pass filter to remove the noises introduced by Laplacian operation. Fig. 1 illustrates the problem by testing an existing text detection method [14] that uses wavelet and color features, in which automatic parameter tuning is also applied for binarization [18] before Laplacian and after Laplacian with the help of Optical Character Recognizer (OCR) available publicly [19]. It is noticed from Fig. 1 that for the images shown in (a), the text detection method detects texts but gives more false positives for the input image compared to the enhanced image. On the other hand, the same text detection method misses a few texts and gives more false positives for the Laplacian image compared to the input image due to noise effect. That is, the text detection method detects texts properly and gives fewer false positives for the enhanced image given by the proposed model (it will be discussed later in proposed methodology section) compared to both of the input and the Laplacian images. The same conclusion can be drawn from the results shown in Fig. 1(b)–(d), where the OCR engine misses a few characters for the texts in the input image and gives garbage values for the texts in the Laplacian image, but correctly recognizes the texts in the enhanced image. This shows that there is a need for a generalized model to remove such operational effects.

From the above discussion, we can infer that there is no consistent enhancement method for reducing the noise effect of Laplacian operation. To the best of our knowledge, there is no generalized enhancement model for the distortions caused by Laplacian or gradient operations so far.

2. Related work

Several methods have been developed for text detection and recognition in video in the past decade [1,2]. We can classify the existing methods broadly into (1) Connected component based, (2) Texture based, and (3) Gradient and edge based methods. Connected component based methods are simple, which explore the properties of character components for text detection. Since these methods use the characteristics of character components, shape analysis of characters is always required. However, due to the distortion effect caused by Laplacian operation as well as low resolution and complex background of video, it is hard to get accurate shapes of character components. Therefore, these methods may not give promising results for video text detection. For example, Rong et al. [20] proposed a two level algorithm for text detection in natural scene images based on the characteristics of character components. Chen et al. [21] proposed a method for robust text detection in natural scene images using Maximally Stable Extremal Regions (MSER) and stroke width distance based features. Yin et al. [22] proposed robust text detection in natural scene images based on MSER, clustering and character classifier. The method studies the characteristics of character components for the output of MSER to classify them as text candidates. Single link clustering and character classifier are used for detecting true text candidates. Since the above discussed methods assumes that a given image has high contrast, the methods like MSER outputs character components. However, if the same methods deployed on video image, the performance of the method degrades severely due to low contrast and low resolution. Therefore, these methods are sensitive to complex background, distortions and low contrast.

To overcome the problem of complex background, many methods have been proposed by using texture features for text detection in video [1,2]. These methods consider the appearance pattern of text as a special texture property. However, since defining the texture property for text components is hard, the methods give a poor accuracy for the texts of different fonts and font sizes. Additionally, most of the methods extract a large number of features and adopts expensive classifiers for improving text detection accuracy. Therefore, they are computationally expensive for real time applications. For example, Shivakumara et al. [14,16,23] proposed wavelet, color features, Fourier with color spaces and Fourier with Lapalacian for text detection in video. These three methods are good for low contrast images but are computationally expensive since they use expensive transformation. In addition, the performances of the methods degrade when an input image contains distortion caused by operations and motions.

To ease the computational burden, methods have been proposed for text detection in video using gradient and edge information. Since edge and gradient information provides significant cues such as high gradient values for text pixels and vital information of the character components lying in vertical and horizontal directions of edges, these methods work well for text detection in video. As a result, they are popular compared to connected component and texture based methods because of simplicity and effectiveness in gradient and edge operations [1,2]. Therefore, most of the state-of-the-art methods explore gradient