Expert Systems with Applications 42 (2015) 7627–7640

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video

Vijeta Khare*, Palaiahnakote Shivakumara, Paramesran Raveendran

University of Malaya, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Available online 6 June 2015

Keywords:
Central moments
Histogram Oriented Moments
Histogram Oriented Gradients
Video text detection
Optical flow
Moving caption text detection

ABSTRACT

Developing an expert text detection system for video indexing and retrieving is a challenging task due to low resolution, complex background, non-illumination and movement of text present in a video. Besides, text detection is vital for several real time applications, such as license plate recognition, assisting a blind person and other surveillance applications. In this paper, we introduce a new descriptor called Histogram Oriented Moments (HOM) for text detection in video, which is invariant to rotation, scaling, font, and font size variations. The HOM finds orientations with the second order geometrical moments for each sliding window (overlapped block) of the input frame. The proposed method performs histogram operations on the orientations of each window to identify the dominant orientation (as a representative). Then, a new hypothesis is defined based on the dominant orientations of a connected component as the numbers of orientations, which point towards centroid of the connected components are larger than the number of dominant orientations which point away from the centroid of the connected components. The components that satisfy the above hypothesis are considered as text candidates, or else as non-text candidates. Further, to detect a moving text- we explore optical flow properties, such as velocity of text candidates to estimate the motions between temporal frames. The components which move with constant velocity and uniform direction are considered as text candidates otherwise non-text candidates. We demonstrate the proposed method's dominance over state of the art methods by testing on benchmark database, namely ICDAR 2013 and our own video datasets in terms of recall, precision and F-measure.

1. Introduction

In the last few years, advancement of new technologies in the field of information retrieval has been changing in day to day life of humans (Jung, Kim, & Jain, 2004; Sharma, Pal, Blumenstein, & Tan, 2012). It is evident from the official statistics of the popular video portal, YouTube, that almost 60 h of videos are uploaded every minute and more than 3 billion videos are watched per day over YouTube. Therefore, retrieval of the videos on World Wide Web (WWW) has become a very important and challenging task for researchers. Secondly, for such a huge database, the conventional content based image retrieval methods may not give an efficient and accurate solution due to the gap between low level and high level features (Jung et al., 2004; Sharma, Pal, & Blumenstein, 2012). To overcome this problem, text detection in video or images has been introduced. It helps to find the meaning, which is close to content of the video/images with the help of Optical Character Recognition (OCR) engines (Fernandez-Caballero, Lopez, & Castillo, 2012; Grafmuller & Beyerer, 2013; Park & Kim, 2013). Thirdly, text detection and recognition in video can be used in several real time applications, such as assisting blind person, assisting intelligent driving, assisting tourists to spot place with the help of GPS, tracking vehicles based on license plate recognition and other surveillance applications. However, text detection and recognition from video or images are not as simple as text detection and recognition from scanned plane background images because usually video suffers from low resolution, complex background and variations in colors, font, font size, orientations and text movements (Chen & Odobez, 2005; Jung et al., 2004; Liu, Wang, & Dai, 2005; Shivakumara, Phan, & Tan, 2010; Wei & Lin, 2012). Therefore, the traditional document analysis methods may not be suitable for text detection in video or natural scene images because these methods require complete shape of the characters and high resolution images. Generally, video contains two types of texts, namely, (i) graphics/superimposed texts, which are inserted by an editor and (ii) scene text, which is part of the image embedded in background. Since graphics text are edited text, they have good clarity and

* Corresponding author at: C2A, Block-C, Residential College 12, University of Malaya, Kuala Lumpur, Malaysia. Tel.: +60 1112282697.

E-mail addresses: kharevijeta@gmail.com (V. Khare), hudempsk@yahoo.com (P. Shivakumara), ravee58@gmail.com (P. Raveendran).

http://dx.doi.org/10.1016/j.eswa.2015.06.002
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

visibility. They are easy to process while a scene text naturally exist and the characteristics are unpredictable so that it is hard to process (Chen & Odobez, 2005; Jung et al., 2004; Liu et al., 2005; Shivakumara et al., 2010; Wei & Lin, 2012). The presence of both graphics and scene texts in video increases the complexity of the problem. As noted (Risnumawan, Shivakumara, Chan, & Tan, 2014) text detection and recognition are not new problems for the document analysis field. The same document analysis based techniques are also extended to solve the problem of text detection from natural scene images (Pan, Hou, & Liu 2008; Pan, Hou, & Liu, 2011; Risnumawan et al., 2014). However, these methods require high resolution and still images but not video because generally natural scene images are captured through high resolution cameras while videos capture through low resolution cameras. Therefore, the methods may not be used directly for text detection in video. We therefore propose a new descriptor called Histogram Oriented Moments (HOM) to overcome the above problems by detecting static and moving text detection in video.

2. Related work

A large number of methods have been proposed in the literature for detecting text in video. These can be classified into two broad categories (Wang & Chen, 2006), (1) the methods which do not use temporal information and (2) the methods which use temporal information. The methods fall on category-1 generally use first frame or key frame of video for text detection. These methods either assume key frames containing text is available or use existing methods for extracting key frames. The methods that fall under category-2 prefer to use temporal information for enhancing text of low resolution or reducing false positives but not for tracking text or for the detection of moving text.

Category 1 can be classified further into three classes, namely, connected component based methods, which exploit characteristics of text components for text detection in video because text components properties help us to separate background from the text information. For example, several methods are discussed in the survey by Jung et al. (2004) where we can notice that the methods proposed geometrical features of text components for segmenting text in video, as well as images. In addition, color has been used for detecting text in video based on the fact that character components in a text line have a uniform color. Chen and Odobez (2005) proposed a sequential Monte Carlo based method for text detection in video. This method uses Otsu thresholds to segment initial text regions and then it uses distribution of pixels of each segmented region for classification of text pixels from the background. Liu, Song, Zhang, and Meng (2013) proposed a method for multi-oriented Chinese text extraction in video. This method uses the combination of wavelet and color domains to obtain text candidates for the given video image. For each text candidate, the method extracts features at component level for classifying component as text or non-text. It is observed from the discussion on connected component based methods that these methods focus on caption or superimposed text but not scene text because caption text has better quality and contrast compared to its background. As a result, the methods expect the shape to be preserved as in document analysis and hence, these methods use uniform color features and shape features. Therefore, these methods are sensitive to complex background because components in background may produce text like features. In addition, these methods are limited to high contrast text but not to scene texts which can have variation in contrasts.

To overcome the problems associated with connected component based methods, texture based methods have been proposed for text detection in video which considers appearance of a text pattern as a special texture. For example, Shivakumara et al. (2010) proposed a method based on the combination of wavelet and color features for detecting text candidates with the help of k-means clustering. Boundary growing has been proposed to extract text lines of different orientations in video. Wang and Chen (2006) have proposed spatial–temporal wavelet transform to enhance the video frames. For the enhanced frame, this method extracts a set of statistical features by performing sliding window over an enhanced image. Then a classifier was used for classifying the text and non-text pixels. Anthimopoulos and Gatos (2013) proposed a method for artificial and scene text detection in images and videos using a Random forest classifier and a multi-level adaptive color edge local binary pattern. The multi-level adaptive color edge local binary pattern has been used to study the spatial distribution of color edges in multiple adaptive levels of contrasts. In continuation, gradient based algorithm has been applied to achieve text detection in video/images. It is noted from the review of texture based methods that most of the methods use a large number of features and classifier with a large number of training samples. Therefore, these methods are said to be computationally expensive though the methods work well for complex background in contrast to connected component based methods. In addition, the methods scope is limited to use with specific languages because of constraints of classifiers and training samples.

To alleviate this problem, gradient based methods have been proposed for text detection in video. These methods work based on the fact that text pixels exhibit high contrast compared to image background and spatial relationship between strokes provide unique properties to differentiate text from non-text. For example, Liu et al. (2005) extract a set of statistical features from the edge images of different directions. Then k-means clustering has been used for classifying text and non-text pixels. Geometrical properties have been used for grouping text pixels and to extract text line in video and images. Wei and Lin (2012) proposed a robust video text detection approach using SVM. This method generates two downsized images for the input image and then performs gradient difference for the three images including the input image which results in three gradient difference images. K-means clustering is applied on the difference images to separate text cluster from non-text. Finally, the SVM classifier has been used for classifying true text pixels from the text clusters. Shivakumara, Phan, and Tan (2009) derive rules using the different edge maps of the input image. The rules have been used for segmenting text region and then the same rules are modified for extracting text information from the video images. Lienhart and Wernicke (2002) have proposed a method based on the combination of gradient and RGB color space. This results in different directions of edge maps for the input image. Then neural network classifier is applied for separating text and non-text pixels. Further, refinement has been proposed for full text line extraction. Zhang and Kasturi (2014) proposed a text detection method based on character and link energies. The method explores stroke width distance to define link energies between the character components on the basis of stroke width distance of the character components is generally almost same. Then a maximum spanning tree is used for text line extraction from both images and videos. It can be inferred from the literature review on edge and gradient based methods that these methods are fast compared to texture based methods, but these are sensitive to background because edge and gradient are not robust to background variations. This results in more false positives.

Overall, it can be concluded from the above discussions that most of the methods in use utilize still images or individual frames extracted from video for text detection. Besides, the main objective of these methods is to detect static text in the images and videos but not moving text in video. As a result, these methods do not