International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137

Volume – 02, Issue – 04, April – 2017, PP – 15-20

ROBUST TEXT AND FACE DETECTION FROM VIDEO

M.NANDHITHA MEENA[1], E.DIVYA BHARATHI[2],
P.JOHN THANGAVEL[3]

[1,2]*B.E Students, Department of ECE, Jeppiaar SRR Engineering College, Chennai,India*
[3]*Asst.Professor, Department of ECE, Jeppiaar SRR Engineering College, Chennai, India*

Abstract: The aim of this paper is to detect the text and face from any video, including the number plate extraction. Text in images contains valuable information and is exploited in many content-based images and video applications. Text detection involves stages like finding the location of text, removing the noise from the text region, extracting the characters, separating the text region from the non-text region, andfinally detecting the text. The role of text detection is to find the image regions containing only text that can be directly highlighted to the user or fed into an optical character reader (OCR) module for recognition. In the fast moving vehicles, a person's face can be detected and extracted from the given video. This work can be useful for camera based security purposes. The proposed text and face detection is implemented using the Matlab software.

Keywords: detection, extraction, OCR module, video, Matlab

1. Introduction

Digital image processing is the process using computer algorithms to perform Image processing on digital images.It allows a wider range of algorithms to be applied to input data and can avoid problems such as a buildup of noise and signal distortion during processing. Of all the applications of image processing,this paper,text and face detection from video comes under video processing application. There is increasing demand for information categorization and retrieval of information, hence some effort has to be taken for detecting text and face from an image or video. Image and video may contain text and face, which helps us extracting useful information. Text and face present in a video consist of a lot of knowledge that belongs to the multimedia system. There are so many applications such as document processing,image indexing, video retrieval and video content summary, in which a very common problem is to extract the text and face from any video. There are three types of text: document text, caption text, and scene text. Document isobtained by scanning printed documents, journals, handwritten historical document and book cover etc.Caption text is usually superimposed on the image.Scene text is a type of text that accidentally appears on the image. It contains text, which is present in the natural part of the scene and contains very important linguistics knowledge liketext on the vehicle, street signs,bill boards, banners so on. Scene text is usually affected by camera parameters such as illumination, focus,arbitrary text layouts, multi scripts, artistic fonts, color, etc.

2. Related Works

The Maximum Stable Extremal Regions methods are giving the better performance in real time projects. But current MSER based methods are still having some drawbacks. For example, they will suffer from detecting of insufficient text candidates construction and repeating components algorithms. In this paper, we will review the MSER based methods focusing on these two problems. Other scene text detection methods can be referred to some survey papers [1]. The algorithm to detect the video text for low and high contrast images [2]. The low and high contrasts are classified by calculating the edge difference between sobel and canny edge detectors. After the process of calculating edge and texture features high contrast and low contrast thresholds are used to extract text region from low and high contrast images separately. The inner distance based shape filter and an intensity histogram based filters are used to eliminate the false positives and extract the text region whose intensities are similar to those of their components coming from the same object and the adjoining areas of the object [3]. Ujjwal Bhattacharya and bityutpropose a work, automatic detection of character components from video frames of class room lectures [7]. There exist a few works on detecting the text from the white boards from video frames. The work of text localization does not exist on black, blue, green and etc. And then the features along with perception network had not reported earlier.

The video frames are used in the present study, which had been collected from video lectures available in online and thus the results of this experimentation are reflecting the robustness of the proposed method. But this method will work properly detect the text on the white board. Hyung Koo and Duck Hoon Kim [8] present a novel scene text detection algorithm on the machine learning technique. Here this technique uses two classifiers.

One classifier is used to generate the character candidates. And another classifier is used to filter the non-text candidates. So the symbols and numerals are not identified by this method.

3. Problem Description

Although many research works have been carried for detecting text and face, there is no proper system for detecting both text and face region from the video. Those systems fail to recognize them when the image or video frame is available in low quality, which may contain valuable information. The proposed framework is able to effectively detect text strings in arbitrary locations, sizes, orientations, font, colors and variations of illuminations or shape of attachment surface.

4. Proposed System

Text and face detection is an important application of image processing. The video is a visual multimedia asource that combines the sequence of images to form a moving picture. The quality of the image depends on the number of frames per minute. The video will be converted into multiple frames. Of which, two frames are selected, one for text detection and other for face detection. The proposed system uses Maximally stable extremal regions(MSER) algorithm for text detection and Viola jones haar cascade algorithm for face detection.MSER is used as a method of blob detection in images. This technique was proposed by Matas et al.to find out the correspondences between image elements from two images with different viewpoints. This method of extracting a comprehensive number of corresponding image components contributes to the wide-baseline matching, and it has led to better stereo matching and object recognition algorithms.The Viola–Jones object detection framework is the first object detectionmethod to provide competitive object detection rates in real-timeand Michael Jones.Although it can be trained to detect avariety of object classes, it was motivated primarily by the problem of facedetection.Fig.3. shows the step by step process of text detection, where the input was given as a frame from the video which contains the text.

5. Text Detection From Video

5.1Gray Scale Conversion

The initial process involved in the proposed system is grayscale conversion. When the input was given as a color image it will be converted into gray scale image, which results in 0's 1's in the matrix when realized using the Matlab software.When the input itself is a grayscale image, then it skips to the next step.A grayscale or grayscaledigital image is an image in which each pixel value is a single sample, that is, it carries only intensity information. Images of this sort, also known as black-and-white, are composed exclusively of shades of gray, varying from black having the weakest intensity to white having the strongest. Grayscale images have many shades of gray in between. Grayscale images are often the result of measuring the intensity of light at every pixel in a single band of the electromagnetic spectrum (e.g. infrared, visible light, ultraviolet, etc.), and in such cases, they are monochromatic proper when only a given frequency is captured. But also they can be synthesized from a full color image; Conversion of a color image to grayscale is not unique; different weighting of the color channels represent the effect of shooting inthe black-and-white film with different colored photographic filters on the cameras. In figure 1, the image is already a grayscale image, hence it will be taken as the output for the grayscale conversion process.

Fig. 1. Input frame

Fig. 2. Grayscale image

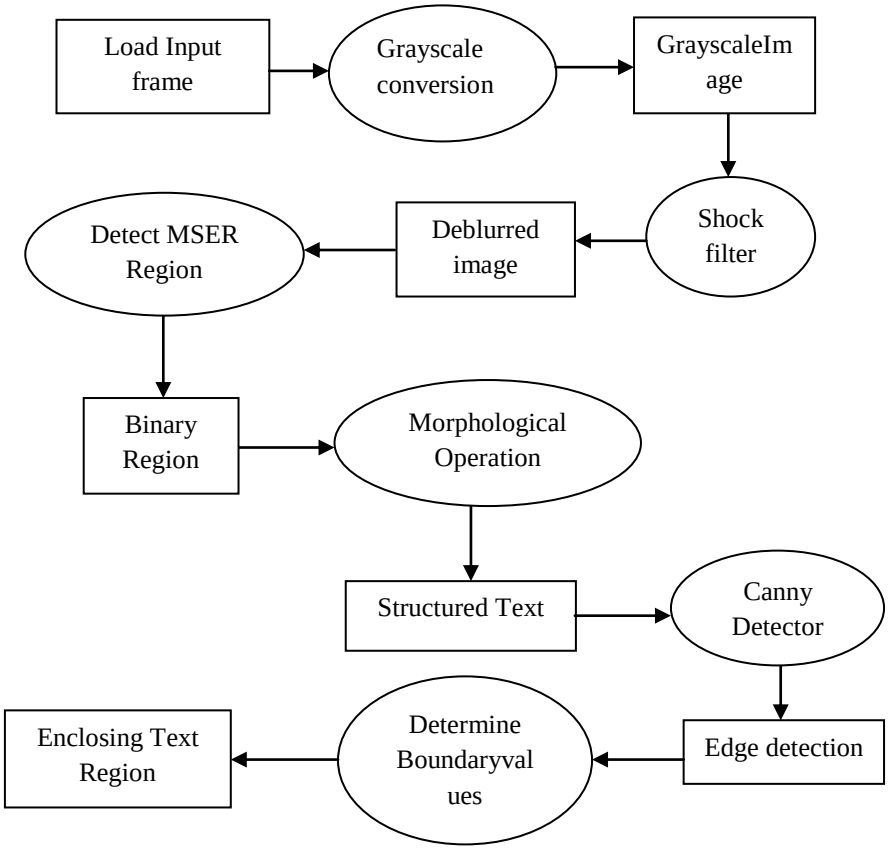

Fig. 3. Dataflow diagram

5.2 Filtering

Image filtering is utilized by many applications, including smoothing, sharpening, removing noise, and edge detection. A filter is defined by a kernel, which is a small array applied to all pixels and its neighbors within an image. In most applications, the center of the kernel is aligned with the current pixel and is a square with an odd number (3, 5, 7, etc.) of elements in each dimension. The process used for filtering an image is known as convolution and may be applied in either the spatial or frequency domain. In this paper, we have shock filter for image deblurring. It creates a "shock" between two influencezones, one belonging to a maximum and the other to aminimum of the signal.By using this PDE (Partial Differential Equation) process,a piecewise constantsegmentation of the input image can be obtained, thus leading to a deblurredoutput.Another advantage of shock filtersover other image enhancement methods is that phenomena like the Gibbsphenomenon cannot appear.

Fig. 4. Deblurred image

5.3 Character Extraction

MSER Algorithm belongs to the existing method called as the connected component method, but here it realized as character candidates rather than connected components.Before extracting the characters, thestable region of the given image is found as shown in fig. 5. It is based on the intensity and the outer borders. It

removes the repeating components which become the major problem for the next step grouping.An extremal region is a connectedcomponent of an image whose pixels have eitherhigher or lower intensity than its outer boundary pixels. The characters are extracted using parent-child relationship as shown in fig. 6.It is very safe and prevents all thecharacters after elimination when the MSER tree is pruned by applying this type of parent children elimination againand again.

Fig. 5. MSER region

Fig. 6. Character extraction

5.4 Text Construction

The text is constructed by using single link clustering algorithm as shown in Fig. 7. Initially, it produces an elongated cluster, which is used for text construction process. Each character is treated as a singleton cluster, then clusters are merged to form a final cluster. In this method, two clusters having the smallest distance are merged in every step. Whenever the distance between the neighborhood clusters exceeds the given threshold value, the clustering process is terminated. The clustering algorithm is based on a good distance metric. Distance metric learning learns the distance function by minimizing the distance between pairs of points in different clusters while maximizing the distance between pairs of points in the same cluster.

Fig. 7. Text construction

5.5 Non-Text Elements Elimination

MSER Algorithm detects many other stable regions in the image that are non-text. A character classifier is used to determine the posterior probabilities of text candidates correspondingto non-text and remove text candidates with high non-textprobabilities. This character classifier will reject the non text elements among the normalized images. The aspect ratio for these images is split into patches of letters and all the characters are separated. These separated characters can be compared with trained data blocks. The matched letter are declared text and each letter blocks can be concatenated to give a word and line as shown in fig. 8.

Fig. 8. Non-text elements elimination

5.6 Text Detection

After all the above steps, the forward-backward algorithm is used to detect the text.(char, char) the pair is at the highest priority rank, (non-char, char) pair is at the medium priority rank and (non-char, non-char) pair is at the lowest priority rank. Forward- backward algorithm begins with the component pair having the highest priority rank. It expands to the forward direction.If the forward direction is no longer expandable, it turns to the opposite direction and tries to expand backward. After the backward direction is no longer expandable, finished the detection of one text lineas shown in Fig. 9. Thus, it detects the text from all other lines. To be qualified as a candidate, a component should meet the angle requirement and the distance requirement.

Fig. 9. Detected text

6. Face Detection From Video

Viola jones haar cascade algorithm is used to detect the face from the given input frame of the video. The main goal of this algorithm is to distinguish the face from the non-faces. Rather than scaling the whole image, it scales the features, which makes it more effective. It has four stages:Haar Feature Selection, Creating an Integral Image, Adaboost Training, Cascading Classifiers.The features that are sought by the detection framework universally involve the sums of image pixels within rectangular areas. As such, they bear some resemblance to Haar basis functions, which have been previously used in the image-based object detection.However, since the features used by Viola and Jones all rely on more than one rectangular area, they are generally more complex. The value of any given feature is the sum of the pixels within clear rectangles subtracted from the sum of the pixels within shaded rectangles. Rectangular features of this sort are compared to alternatives such as steerable filters. Although they are sensitive to vertical and horizontal features, their feedback is considerably coarser.

Haar Features – All human faces have some similar properties. These properties may be matched using Haar Features. A few properties common to human faces:

- The eye region is darker than the upper cheeks.
- The nose region is brighter than the eyes.

The composition of properties forming matchable facial features:

- Location and size: eyes, mouth, bridge of nose
- Value: oriented gradients of pixel intensities

6.1 Rectangle features

- Value = Σ (pixels in black area) - Σ (pixels in white area).
- Three types: two-, three-, four-rectangles, Viola & Jones used two-rectangle Features.
- For example: In brightness,the difference between the white &black rectangles over a specific area.
- Each feature is related to a special location in the sub-window.

6.2 Integral image

An integral image is an image representation thatevaluates rectangular features in constant time, which gives them a considerable speed advantage over more sophisticated alternative features. The rectangular area of each feature is always adjacent to at least one other rectangle, it follows that any two-rectangle feature can be computed in six array references, any three-rectangle feature can be computed in eight, and any four rectangle feature can be computed in nine. The integral image at location (x, y), is the sum of the pixels above and to the left of (x, y), inclusive.

In Fig(11), the face has been detected from the given image. .After detecting the face, shock filter is used to deblur the face region or removes the noise from the face region as shown in Fig. 12.

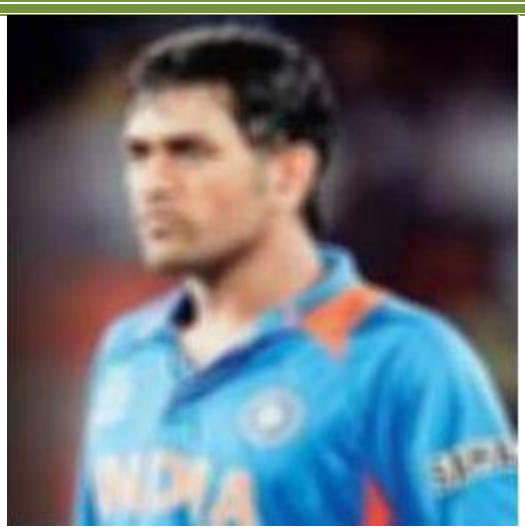

Fig. 10. Input frame

Fig. 11. Detected Face

Fig. 12. Deblurred image

7. Conclusion

In this paper, we have proposed a system that can detect the text and face present in the video, which will be useful for camera based security problems.We have presented an improved text detection method which can effectively detect text from the complex background. Due to the unpredictable text appearances and complex backgrounds, text detection in natural scene images is still an unsolved problem to locate text regions present in those images.Thus, the proposed framework is a novel technique that can detect both text and faceand works with the images of low quality, images of different sizes and different shapes with less computation cost.

References

[1]. K. Jung, K. Kim, and A. Jain, (2004) "Text Information Extraction in Images and Video: A Survey" International Conference on Pattern Recognition, Vol 37, No. 5, pp 977-997.

[2]. Shivakumara, p., Huang, W., Phan, T.Tan, C.L. (2010) "Accurate video text detection through classification of low and high contrast images", Pattern Recognition Vol. 12, No. 3, pp 738- 750.

[3]. Liu, Z., Sarkar, S, (2008) "Robust Outdoor Text Detection Using the Intensity and Shape Features", International Journal of Pattern Recognition,Vol. 14, No. 5, pp 675-693.

[4]. X. Yin, X-C. Yin, H.W. Hao, and K. Iqbal, (2012) "Effective Text Localization in Natural Scene Images with MSER, Geometry Based Groupingand Adaboost," in Proceedings of International Conference on Pattern Recognition, Vol. 21, No. 9, pp. 4256– 4268.

[5]. J. Dina, M. Vanitha, S. Pradeepkumar, B. Udhaya, and Saravana A, (2014) "Robust Text Detection and Voice Conversion", in InternationalJournal of Science and Engineering Communications. Vol. 03, No. 3, pp.1062-1068.

[6]. Xu Cheng Yin,Xu Wang Yin., Kaizhu Huang, and Hong Wei-hao, (2014) "Robust Text Detection in Natural Scene Images", IEEE Transactionson Pattern Analysis and Machine Intelligence, Vol. 36, No. 5, pp 143-154.

[7]. Ujjwal Bhattacharya, Bityut B, (2014) "Automatic Detection of Handwritten Texts from Video Frames of Lecturer", IEEE Conference onFrontiers Handwriting Recognition. Vol. 36, No. 5, pp137-144.

[8]. Hyung Koo, and Duck Hoon Kim, (2013) "Scene Text Detection Via Connected Component Clustering and Non-text Filtering", IEEE Journal onImage Processing. Vol. 22, No. 6, pp1057- 7149.

