2009 10th International Conference on Document Analysis and Recognition

Robust Extraction of Text from Camera Images

S P Chowdhury
Intelligent Systems & Control
QUB, UK
schowdhury01@qub.ac.uk

S Dhar
S/W Engineer
Videonetics, India
s.dhar.in@gmail.com

A K Das
CST Dept.
BESU, India
amit@cs.becs.ac.in

B Chanda
ECS Unit.
ISI, India
chanda@isical.ac.in

K McMenemy
Intelligent Systems & Control
QUB, UK
k.mcmenemy@ee.qub.ac.uk

Abstract

Text within a camera grabbed image can contain a huge amount of meta data about that scene. Such meta data can be useful for identification, indexing and retrieval purposes. Detection of coloured scene text is a new challenge for all camera based images. Common problems for text extraction from camera based images are the lack of prior knowledge of any kind of text features such as colour, font, size and orientation. In this paper we propose a new algorithm for the extraction of text from an image which can overcome these problems. In addition, problems due to an unconstrained complex background in the scene has also been addressed. Here a new technique is applied to determine the discrete edges around the text boundaries. A novel methodology is also proposed to extract the text exploiting its appearance in terms of colour and spatial distribution.

Keywords: Text extraction, text localization, camera image, video frame, discrete edge boundary.

1. Introduction

Any camera based image can be subject to operations like text information extraction (TIE) for applications such as optical character recognition (OCR), image/video indexing, mobile reading system for visually challenged persons etc. TIE from camera based scene images is a very difficult problem because it is not always possible to precisely define the features of text in a coloured scene image due to the wide variations in possible formats; for example, geometry (location and orientation), colour similarity, font and size. Moreover, camera based images can be subject to numerous possible degradations such as, blur, uneven lighting, low resolution and contrast which makes it more difficult to recognise any text from the background noise. This paper is organised as follows. Section 2 describes relevant past research. The techniques proposed in this paper for text extraction are detailed in section 3 that presents our work module wise in 5 subsections. Section 4 contains experimental results and conclusion.

2. Past Research

Due to its immense potential for commercial applications, research in TIE from camera based colour scene images is being pursued both in academia and industry. There are many reports in the literature such as the survey paper by Jung et al. [5]. TIE techniques can be broadly divided into two classes; i) region based and iii) texture based methods. Region based methods are bottom up approaches where connected components (CC) or edges are found on the basis of their perceptive difference with the background. This is followed by merging of the CCs or edges to get the text bounding boxes. Some of the CCs and edge based methods are described in [2, 3, 6, 8, 11]. The basic understanding for texture based methods is that typically text in an image has distinct textural properties that distinguish it from the background. Some of the representative methods are given in [1, 4, 7, 10]. Different tools used for texture analysis include the Gabor filter, Fast Fourier Transform and wavelets.

3. Proposed Work

In this research, we aim to exploit some of the properties of text which is common among different languages. The major advantage of our work is that it is OCR independent. Written text has generally some distinct property that can help humans to correctly locate it on camera based images (still or video). Even lack of familiarity with the alphabets for different languages does not create problem for this recognition. Few of properties are listed.

1. Easily distinguishable text color that is significantly different from background.
2. Boundaries of the characters are smooth, that is, there are not many protrusions.
3. For a character, in a single row scan there exits at least one color band covering the characters stroke-width.

These three observations have led us to develop a new technique for extraction of text from an image. A detailed description of how to segment the color chains belonging to

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.188

text within an image is given in section 3.1. As already mentioned, in some cases the edges of the text within the images may be blurred. To enhance such edges locally an edge enhancement technique was used. This is explained in section 3.2. Typically, edges would be "de-blurred" before the analysis detailed in section 3.1 carried out.

The unique approach of our proposed technique is to correlate together the color based segmentation methodology and the spatial distribution based pattern findings which is not prominently reported in any other research. Individually, color segmentation and spatial distribution based pattern findings are two consecutive processes and the later one is carried out once it is fully segmented from the color domain to binary domain. In our approach we have merged these two and either segmented the color image in only the horizontal direction and matched the pattern of this segmented portion (section 3.3) in the vertical direction or viceversa. Following this, it is then possible to regroup probable text blocks, this is detailed in section 3.4. This is then followed by the removal of non text block in section 3.5.

3.1. Color chain segmentation

To link consecutive pixels in the same direction, color chains are segmented from that image by exploiting the general text properties described earlier. Let us define the color image F as a two dimensional function of three color values. In each dimension, the color value can range form 0 to $L-1$. $F^R(i,j)$, $F^G(i,j)$ and $F^B(i,j)$ represents the red, green and blue color component of i^{th} row and j^{th} column position. The color euclidean distance of two coordinates (i_1,j_1) and (i_2,j_2) abbreviated by a and b is $E(a,b)$;

$$E(a,b) = \sqrt{[(F^R(a)-F^R(b))^2+(F^G(a)-F^G(b))^2+(F^B(a)-F^B(b))^2]}.$$

For segmenting the color chains, the threshold values for checking color similarity is kept at 30% in each color plane and the euclidean distance among two color values is kept at 45% of the farthest possible euclidean distance $(\sqrt{3}L)$. Now consider two consecutive pixels (i_1,j_1) and (i_1,j_{1+1}) along the same row. These two pixels will be assigned to the same chain if the following conditions are satisfied.

$$|F^X(i_1,j_1)-F^X(i_1,j_1+1)| < 0.3L$$
$$where, (X=R,G,B) \ and$$
$$E((i_1,j_1),(i_1,j_1+1)) < 0.45(\sqrt{3}L)$$

An example of visual representation of the color chains is shown in the Figure 1(c). Consecutive color chains in the same row have been labelled with alternative black and white runs. Figure 1(a) illustrates the original image with a marked region. Figure 1(b) and (c) are the magnified images of the marked region and color chains into that region.

(a) (b) (c)

Figure 1. Color chain segmentation.

3.2. Edge enhancement technique

In the process of color chain segmentation some difficulties may be encountered because of smooth transitions at the edge of character components e.g. due to image blurring. To overcome this, were used preprocessing operations on the input image to obtain sharp transitions at the edges. Popular edge enhancement techniques like the sobel filter enhances the edge using the intensity information of its eight neighbor pixel values. In our case this is not sufficient as edges are continuously changing with a small gradient. The stabilized inverse diffusion equation is used by Pollak et al [9] to enhance the edges. This algorithm is good for images where the number of different regions are known a priory, which is also not applicable in this research.

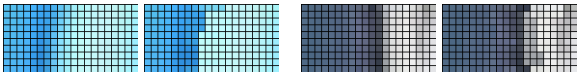
A two pass enhancement technique is proposed for the enhancement of the edges for our new algorithm. In the first pass, it enhances all the fluctuating but prominent gradient transitions. This is known as light edge enhancement (LEE) and the objective is to find a set of consecutive candidate points and enhance the edge between them. LEE needs to be operated in the same direction that we want to segment the color image. In LEE, the first step is to obtain the set of candidate points thus, it is necessary to consider all points in the raster scan direction.

Let us define the horizontal gradient value function GV^X and horizontal gradient sign function GS^X on a particular color plane X. We may write;

$$GV_H^X(i,j) = F^X(i,j) - F^X(i,j+1)$$
$$GV_H^X(i,j) < 0 \Rightarrow GS_H^X(i,j) = -1$$
$$GV_H^X(i,j) = 0 \Rightarrow GS_H^X(i,j) = 0$$
$$GV_H^X(i,j) > 0 \Rightarrow GS_H^X(i,j) = 1$$
$$where, (X=R,G,B)$$

To obtain a set of candidate points, start from a position (i,j) and store the initial gradient sign value in S^R, S^G, S^B. Continue to check the gradient until there is no other point (k,l) consist of at least one color plane Y, for which $|F^Y(i,j)-F^Y(k,l)| \geq 0.3\%L$. At any point (p,q) the gradient checking condition for continuing accumulation in the candidate points is $GS^X(p,q) = S^X \ \ or \ \ |GV^X(p,q)| < 0.05L$ where $(X = R,G,B)$. Once this condition is not fulfilled, LEE stops and starts a new search from the next

point. On availability of a set of successful and consecutive candidate points we find the point which is in the middle of the total color distribution among the candidate points in the Y plane which has the maximum change in color value between the maximum and minimum points. This middle point will come within the maximum and minimum color valued points in Y plane. Using this point as a gate we change all the color values in minimum side with the minimum value and with the maximum value in maximum side. Thus the edge information is significantly enhanced. Figure 2(a) describes where LEE is required and Figure 2(b) illustrates the outcome of LEE.



(a) (b) (c) (d)

Figure 2. LEE and HEE.

Next, on the output of LEE, heavy edge enhancement (HEE) will be carried out. The main difference between HEE and LEE is that it needs a strictly positive or negative gradient throughout the candidate pixels for all three color planes. The condition for accumulating the candidate points for HEE is $GS^X(p,q) = S^X$ where $(X = R, G, B)$. After finding the candidate points, enhancement of the edge will be achieved in same way as it was for LEE. It is also necessary to find the color mean point among the color distribution in the candidate points. The purpose of HEE is to find the consecutive points where the gradient is strictly of same sign for a color plane. All three color planes need to follow this condition. Figure 2(c) describes where HEE is required and Figure 2(d) is the outcome after HEE.

These edge enhancement techniques are used as a preprocessing step on the camera images. An added advantage of this edge enhancement technique is that it is also useful to remove or reduce motion blur from camera image sequences. That allows us to treat images from the still camera and frames from the video images in the same manner.

3.3. Spatial distribution based pattern finding

After the color chain segmentation is performed on the enhanced edge image, the next step is to link the chains vertically to get a two dimensional spatial color component. A number of features that have observed on the vertical distribution of the horizontal chains for a particular character are listed.

1. In a single row there exits at least one color chain.
2. Horizontal chains cover the horizontal continuity of the characters.
3. The size of the chains that are positioned vertically one after another, has mainly three types of relation for different reasons such as:
   (a) Almost of same size; It is because of the same stroke width of a characters or same length of a protrusion.
   (b) Continuous change in size; The main reason behind this is circular bending in the character.
   (c) Sudden change in size; It only happens when some protrusions that come out of the character
4. The average color values of two chains that are positioned one after another vertically are almost the same unless written with a high contrast color mixtures; this is not very common.

Based on these observations, horizontal chains are linked vertically to form a two dimensional object. Let us number the chains chronologically 1 to N and define three functions LC, RC and WC to define the left most, right most column positions and width of a chain. Average red, green and blue color values of the k^{th} $(1 \leq k \leq N)$ chain is denoted as $A^R(k)$, $A^G(k)$ and $A^B(k)$. Function MIN selects the minimum value among a set of values. For two vertical neighbor chains a and b $(1 \leq a, b \leq N)$, presence of vertical overlap can be established by either of $(LC(a) \leq LC(b) \leq RC(a))$ or $(LC(b) \leq LC(a) \leq RC(b))$. If there is an overlap at all then let the amount of vertical overlap of a and b chains be calculated by a function OV where,

$$OV(a,b) = MIN((RC(a) - LC(b)), (RC(b) - LC(a)))$$

and if there is no vertical overlap among a and b then $OV(a,b) = 0$. The euclidean color distance function of the average color of two chains (a and b) is defined as EC; where

$$EC(a,b) = \sqrt{[(A^R(a) - A^R(b))^2 + (A^G(a) - A^G(b))^2 + (A^B(a) - A^B(b))^2]}.$$

To get a set of groups representing the characters fully or partially, we have imposed two sets of conditions to link the selected chains vertically. Selection of the chains are made on the basis of their widths. Chains having width more than the width of biggest possible character (WBC) is rejected. For any selected chain c $(1 \leq c \leq N)$, it is necessary that $WC(c) \leq WBC$. This is to remove noise from the output and it is justified as this whole process is carried out without having any resolution information. Here we mention two sets of conditions to satisfy the observations 3a and 3b mentioned in this subsection. If a and b $(1 \leq a, b \leq N)$ have to satisfy the first condition then there has to be a significant vertical overlap between them; such that

$$OV(a,b) > 0.8MIN(WC(a), WC(b)).$$

The second set of conditions is to allow the flattened part of the characters to be identified, situation like bending. Here, in addition to the first condition with more relaxation, we have used the average color distance among the chains to exploit the observation number 4.

$$\begin{aligned} OV(a,b) > 0.5MIN(WC(a), WC(b)) \quad & and \\ |A^X(a) - A^X(b)| < 0.3L \ \ where, (X = R, G, B) \quad & and \\ EC(a,b) < 0.45(\sqrt{3}L) & \end{aligned}$$

The output from this process is a set of candidate groups that contains vertically linked chains. Now restriction of the candidature is the number of chains present in that group. All the groups having a number of chains less than the height of the smallest possible character (HSC) are ignored. Let us assume that there are M number of groups formed which are numbered from 1 to M. Count of the chains present in a group m $(1 \leq m \leq M)$ is presented by a function $GC(m)$. Thus the criterion to be a candidate group is $GC(m) \geq HSC$.

3.4. Regrouping of the probable text blocks

In this part of work, all the candidate groups are put in a two dimensional matrix I having same size as the input image. For any position (i, j), if it is part of any chain belonging to a candidate group then we put $I(i, j) = 1$ else $I(i, j) = 0$. Now using the methods described in subsections 3.1 and 3.2 vertically instead of horizontally a new set of P number of vertical chains are obtained. Let us define two function TC and BC on these vertical chains such that for any vertical chain p $(1 \leq p \leq P)$, $TC(p)$ will give the top most position and $BC(p)$ will give the bottom most position of the p^{th} chain.

Now with a vertical ortho-raster scan on I we find all the gaps on the series of $1s$ or all the protrusions on the series of $0s$ having length less than HSC and not shielded by $TC(p)$ or $BC(q)$. All the gaps of 0 are filled with 1 and all the protrusions are removed by 0. This step is important to remove miss detection or some false extra detection among the probable text bands.

Initial groups of chains GC are now distorted and we need to regroup these using the matrix I as a template. Here in another matrix J of same size as input image, at any point q, $J(q) = 1$ if,

$$\begin{aligned} LC(s) = q \ or \ RC(s) = q \mid \exists\, s,\ 1 \leq s \leq N \quad or \\ TC(t) = q \ or \ BC(t) = q \mid \exists\, t,\ 1 \leq t \leq P \end{aligned}$$

otherwise $J(q) = 0$. Thus the matrix J will contain all the transition points for both of the horizontal and vertical chains. In Figure 3(a) and Figure 3(b) one original image and its J matrix have been shown. Here original image has a mark region for further description. If there are multiple characters forming a text side by side then there are some additional characteristics for that group of the characters that can be observed such as

1. If the characters are thin and the density of the characters are high then the density of the number of chains in a particular region increases.

2. If there are many characters side by side then there exists repetitions of the chains whose average color value is similar to one another.

3. Spatial characteristics of the chains due to inter character space is almost similar as of characters.

This regrouping will be done using the connectivity obtained from the matrix I and density of transition points (1) present the matrix J. In the new set of groups all the connected points having 1 in matrix I and having similar density of 1 in matrix J will get clubbed. Different densities of the transition points are shown in Figure 3(c) which is the magnified J matrix of the region marked in Figure 3(a). Let us consider that U number of new groups are formed and k^{th} $(1 \leq k \leq U)$ new group has density of transitions points (obtained from matrix J) $TD(k)$. All the groups have some spatial characteristics based on the two dimensional spread of the member points. Denote the height and width of the spread of member points and existence of the k^{th} group as $HG(k)$ and $WG(k)$ and $EF(k)$. In the initial state $EF(k) = 1$ $(\forall k,\ 1 \leq k \leq U)$.

(a) (b) (c)

Figure 3. Transition points and their densities.

Now check for all the groups and remove a group x $(1 \leq x \leq U)$ if $HG(x)$ or $WG(x)$ is small and have lower $TD(x)$ value. This is done by exploiting the observation made on 1 and 3. Removal of a group x is noted by making $EF(k) = 0$. Here the threshold values are obtained easily if we consider that in most of the cases aspect ratio of the character bounding box varies within $0.5\ to\ 2$. And it is obvious that for any of the horizontal or vertical scan line there should be at least two transition points to define the boundary.

Figure 4. Robust text extraction results.

3.5. Removal of non text blocks using repetition feature

Here our aim is to exploit the observation 2 made on subsection 3.4. Towards that goal we use each of the groups x $(1 \leq x \leq U)$ which are the outcome subsection 3.4 and not removed yet $(EF(x) = 1)$. Using the points inside that group as mask, we are recalculating horizontal chains as we have done previously on subsection 3.1 on the edge enhanced image obtained from subsection 3.2. Let us number the chains those come out from a particular group chronologically 1 to Q. Denote average red, green and blue color values of the k^{th} $(1 \leq k \leq Q)$ chain by $A^R(k)$, $A^G(k)$ and $A^B(k)$ as it was done previously in subsection 3.3. Also denote width of the chain and euclidean color distance function of the average color of two chains by WC, EC consecutively. Now in each pair of the horizontal chains a and b $(1 \leq a, b \leq Q)$ come in single row scan, check whether the average color distance follow these similarity conditions,

$$|A^X(a) - A^X(b)| < 0.05L \;\; where, (X = R, G, B) \quad and$$
$$EC(a, b) < 0.07(\sqrt{3}L).$$

For a group x $(1 \leq x \leq U)$ that contains all the chains y which are satisfying the above mentioned condition; add there width $WC(y)$ into CN. Let the sum of the all chains TW $(TW = \sum WC(y)\ \forall y, 1 \leq y \leq Q)$. Now remove the group x $(1 \leq x \leq U)$ for which

$$\frac{HG(x)}{WG(x)} < 0.5 \;\; and \;\; \frac{CN}{TW} < 0.5$$

Removal of a group x will make the value of $EF(x)$ to 0. All the remaining groups y for which $EF(y) = 1$ will be considered for the intended text block of the image.

4. Experimental Results and Conclusion

In order to detect text blocks from a camera grabbed image, we have carried out our experiments using around 270 sample images taken from the "ICDAR'03 Robust Reading" competition data set (http://algoval.essex.ac.uk/icdar/Datasets.html) as well as from our own collection of still images and video frames. A detection rate of text is almost 93% whereas rate of false positive detection is also 4%. Result is computed on the overlap of the output bounding boxes with the ground truth. In most of the cases a logo that is present in an image also will be detected as text. Sometimes text like patterns in background also get detected as text. The text part that was failed to be identified is because of they were faint and difficult to get a prominent gradient transition on their boundary. Extracted text zones out of this methodology are illustrated with the outer marker in Figure 4.

It may be noted that the thresholds used in the algorithm are not very rigid and sensitive. Reasonable variation of the threshold values are possible without appreciable degradation in the final results. Finally, we claim that the novelty of this work is the blending of the color feature and spatial distribution together for text extraction. We have also used a good edge enhancement technique well suited to different types of scene analysis.

References

[1] B. T. C. Y. Bae and T. Y. Kim. Automatic text extraction in digital videos using fft and neural network. In *Proceedings of IEEE International Fuzzy Systems Conference*, volume 2, pages 112–1115, 1999.

[2] Q. L. et al. Text localization based on edge-cca iand svm in images. In *Samsung Techlogy Conference*, 2005.

[3] A. K. Jain and B. Yu. Automatic text location in images and video frames. *Pattern Recognition*, 31(12):2055–2076, 1998.

[4] K. Jung. Neural network based text location in colour images. *Pattern Recognition Letters*, 22 (14):1503–1515, December 2001.

[5] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: A survey. *Pattern Recognition*, 37:977–997, 2004.

[6] C. M. Lee and A. Kankanhalli. Automatic extraction of characters in complex in complex images. *Int. J. of Patter Recognition and AI*, 9(1):67–82, 1995.

[7] H. Li, D. Doermann, and O. Kia. Automatic text detection and traking in digital video. *IEEE Image Processing*, 9 (1):147–155, January 2000.

[8] S. Messelodi and C. M. Modena. Automatic identification and skew estimation of text lines in real scene images. *Pattern Recognition*, 32:791–810, 1992.

[9] I. Pollak, A. S. Wilsky, and H. Krim. Image segmentation and edge enhancement with stabilized inverse diffusion. *IEEE Tr. on IP*, 9(2), February 2000.

[10] V. Wu and R. Manmatha. Textfinder: An automatic system to detect and recognize text in frames. *IEEE Pattern Analysis and Machine Intelligence*, 21 (11):1224–1229, November 1999.

[11] Y. Zhong, K. Karu, and A. K. Jain. Locating texts in complex color images. *Pattern recognition*, 28(10):1523–1535, 1995.