

A New Fourier-Moments based Video Word and Character Extraction Method for Recognition

^aDeepak Rajendran, ^aPalaiahnakote Shivakumara, ^aBolan Su, ^bShijian Lu, ^aChew Lim Tan

^aSchool of Computing, National University of Singapore, Singapore
 {deepak, shiva, subolan and tancl@comp.nus.edu.sg

^bInstitute for Infocomm Research, Singapore, slu@i2r.a-star.edu.sg

Abstract— This paper presents a new method based on Fourier and moments features to extract words and characters from a video text line in any direction for recognition. Unlike existing methods which output the entire text line to the ensuing recognition algorithm, the proposed method obtains each extracted character from the text line as input to the recognition algorithm because the background of a single character is relatively simple compared to the text line and words. Max-Min clustering criterion is introduced to obtain text cluster from the extracted Fourier and moments feature set. Union of the text cluster with Canny operation of the input video text line is proposed to obtain missing text candidates. Then a run length criterion is used for extraction of words. From the words, we propose a new idea for extracting characters from the text candidates of each word image based on the fact that the text height difference at the character boundary column is smaller than that at other columns of the word image. We evaluate the method on a large dataset at three levels namely text line, words and characters in terms of recall, precision and f-measure. In addition to this, we show that the recognition result for the extracted character is better than words and lines. Our experimental set up involves 3527 characters including Chinese. The dataset is selected from TRECVID database of 2005 and 2006.

Keywords- *Video word segmentation, Video character extraction, Fourier-Moments, Run length, Text height difference, Video character recognition*

I. INTRODUCTION

Though lots of content based retrieval algorithms are developed for meeting requirements of video indexing and retrieval in the field of image processing and multimedia, understanding video content and automatic annotations still remain an unsolved problem due to the semantic gap between the high level and the low level features. Therefore automatic extraction of a video text, which aims at integrating advanced optical character recognition (OCR), is vitally useful for video annotation and retrieval systems [1]. Hence video text extraction and recognition is crucial to the research in video indexing and summarization [1-6].

Video text recognition involves four steps: detection, localization, extraction, and recognition. The detection step roughly identifies text regions and non-text regions. The localization step determines the accurate boundaries of the text rows. The extraction step identifies word and character boundaries properly before binarization and recognition [7]. There are several algorithms that are reported in the literature for accurate text detection and they have achieved good accuracy even for scene text detection [8-13] and multi-oriented text detection [13]. Therefore, in this work we use the method that works for multi-

oriented scene text reported in [13] to locate text lines with bounding boxes in video images. It is also true that character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. While OCR systems have been developed for recognizing characters printed on clear paper, applying a current OCR system directly on video text leads to poor recognition rates, typically from 0% to 45% [2]. This is due to the complex background and low resolution of video images. To recognize these video characters, it is necessary to reduce the complex background by segmenting words and characters properly even when the whole text string is already well located. Therefore, the third step is important and research is badly needed to meet requirements of real time applications such as video events analysis and sports events analysis etc. Hence, we focus on word and character extraction from detected text lines in video [1-7].

The text region extraction methods are classified into three classes. Methods in the first class use either global or local or multilevel thresholds to retrieve text regions. The second class uses stroke based methods to retrieve text regions. The third class is the color-based methods. However, performance of these methods is poor because of the complex background and unfavorable characteristics of video images. Recently, a language independent text extraction method [7] is proposed which works based on adaptive, dam point labeling and inward filling. However, this method is sensitive to complex background images.

Chen et. al. [3] proposed a two-step method for text recognition. The first step uses edge information to localize the text and the second step uses features, and machine learning to recognize the segmented text. This method is not robust enough for complex backgrounds. Chen and Odobez [2] proposed a method based on Monte Carlo sampling for segmented text recognition. This method is expensive as it uses probabilistic Bayesian classifier for selecting thresholds. Another method [1] for low resolution video character recognition is proposed based on a holistic approach and connected component analysis. However, it requires a large number of training samples. There are robust binarization methods which take the whole detected text region as input without segmenting a text region into words and characters to improve the recognition rate of video character recognition [14-15]. Recently, Zhou et. al.[15] developed a binarization method for video text recognition. This method uses Canny information to binarize and it achieves a reasonably good accuracy compared to conventional thresholding methods. The above methods focus on graphics text recognition rather than scene text recognition and hence their error rates are high if scene text is present in the image.

It is noted from our review of literature that existing methods accept a whole text region detected by the text detection methods

as input for binarization and recognition. Besides, the methods focus on graphics text and horizontal text. Hence, these methods are not good enough to handle the problems of complex background, multi-oriented text and scene text in the video. However, we have found a method [7] that performs character extraction from the segmented text line based on an assumption that characters have uniform color and that the text is in horizontal direction. These assumptions may not be valid in the case of scene text. Nevertheless, we are inspired by this work to propose a new method for extraction of words and characters before binarization and recognition to achieve better accuracy even for scene text and multi-oriented text in video.

Therefore, in this paper, we propose a novel Fourier and moments based method for word and character extraction from video text lines in any direction. The Fourier transform has the ability to enhance text pixels in video image as it gives high frequency components for text pixels and low frequency components for non-text pixels [16]. To further increase the gap between text and non-text pixels, we propose the use of moments on the inverse transfer of the Fourier image as we know that the moment computation involves intensity and spatial information of the image, to widen the difference between text and non-text pixels [11]. For Fourier-Moments features, we introduce Max-Min clustering to obtain text cluster. The text cluster is combined with the Canny operation on the input text line image through a union operation to obtain missing text candidates. The run length concept is used for word extraction. Character extraction from the text candidate word image is done based on the fact that the text height difference at the character boundary column is smaller than the text height differences at other columns. This will be further explained in Section D.

II. PROPOSED METHODOLOGY

Since our focus is on word and character extraction from video text lines that are detected by the text detection method, we use a method proposed in [13] for video text location with clear bounding box. The reason for choosing this method is that the method is able to locate both graphics and scene text and multi-oriented text in complex video background despite low resolution of video images. In order to ease the problem due to multi-oriented scene text, we take advantage of the angle of text line determined by the text detection method during bounding box fixing. We then use Bresenham's line drawing algorithm [17] to identify the text pixel direction. As a result, we convert text lines of any direction into horizontal text lines. Hence the problem of multi-orientation of text line has been simplified to the problem of horizontal text line.

The proposed method is structured into three subsections. Bresenham's line algorithm for handling multi-oriented text is described in Subsection A. The Fourier and moment combination features are explained in Subsection B to obtain text cluster. In Subsection C, we propose a method to combine text cluster obtained in Subsection B with a Canny operation on the input image to restore the missing text candidates for word extraction. Subsection D presents a method for character extraction from the segmented word image based on a thickness vector, a top distance vector and a bottom distance vector.

A. Bresenham's Line Algorithm

The Bresenham's Line Drawing algorithm [17] is used to determine the points in an n-dimensional raster that should be plotted to form a close approximation to a straight line between two given points. This algorithm works well because it takes the

coordinates of the bounding box determined by the text detection method to compute the direction of line which is then used to convert a text line of any given direction into a horizontal text line as shown in Figure. 1. However, the quality of image degrades somewhat. It can be seen in Figure. 1.

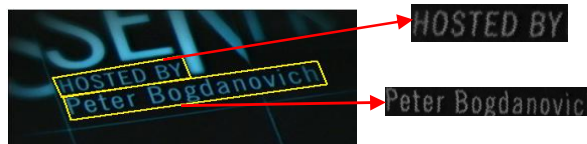


Figure 1. Non-horizontal text line into horizontal text line

B. Fourier-Moments Features

Since video images have low resolution and complex background, we need a mechanism to enhance low contrast text pixels in order to differentiate text from the background. Therefore we propose a Fourier and moments combination as we mentioned before for text enhancement. For a given horizontal text line image, we apply the Fourier transform to get high frequency components for text pixels as the Fourier transform gives high energy for high contrast pixels. Figure 2 shows a gray text image in (a), the Fourier spectra for text pixels in horizontal and vertical directions in (b), and the effect of inverse Fourier transform for the gray image in enhancing the brightness in (c), as compared to the input image. Note that for visualization, we have given the binary form for the spectra of Fourier transform in Figure 2(b)).

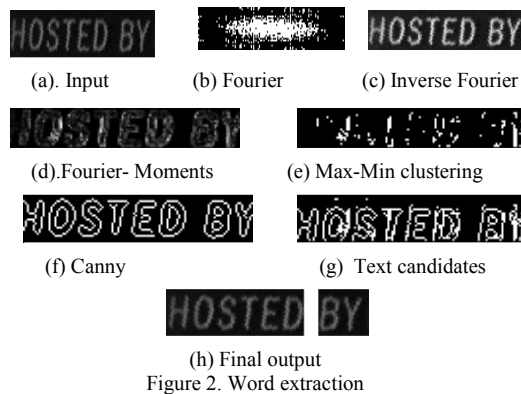


Figure 2. Word extraction

Further, in order to increase the gap between text and non-text pixels, we propose the use of moments on the absolute inverse transfer of Fourier image shown in Figure 2(c). As we know that the moment computation involves intensity and spatial information of the image, which differs between text and non-text pixels. Next, the average of the mean and median moments of the image is computed.

We use a 3×3 sliding window on the resultant image of the absolute inverse Fourier transform (Figure 2(c)) to calculate the moments with respect to mean and median as defined in equations (1) to (6). The average of absolute of mean and median moments is determined as defined in equation (7) for each sliding window for each pixel in the absolute inverse Fourier transform image. As a result, we get a Fourier-moments feature matrix for the text line image. The effect of the average moments over the absolute inverse transform can be seen in Figure 2(d) where one can notice high contrast values at edges of text information. The same thing is illustrated in Figure 3 showing high contrast values at text pixels

and low contrast values at non-text pixels. The gap representing low contrast values for the scan line 15 across the text in Figure 3 is marked by a green color oval. Hence it is confirmed that Fourier-moments combination helps in classification of text and non-text pixels. To classify text pixels from non-text pixels, we introduce the Max-Min clustering criterion instead of determining a threshold value for binarization. The result of Max-Min clustering is shown in Figure 2(e). The Max-Min clustering method selects Max and Min values in the feature matrix and then it compares each value in the feature matrix with Max and Min chosen values to find its nearest neighbor. i.e the value which is close to Max is classified as text and the values which is close to Min classified as non-text. This results in a text cluster as shown in Figure 2(e).

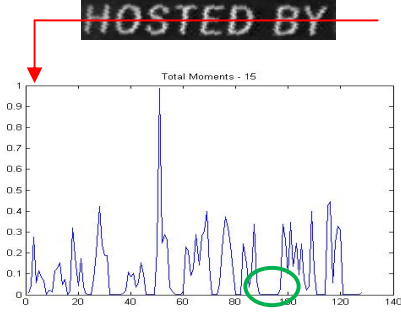


Figure 3. Fourier-moments feature at gaps between the words

1st Order Mean moment

$$M(I) = \text{Mean of the } 3 \times 3 \text{ block.} \quad (1)$$

$$2^{\text{nd}} \text{ Order Mean moment } \mu_2(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (I(i, j) - M(I))^2 \quad (2)$$

3rd Order Mean moment

$$\mu_3(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (I(i, j) - M(I))^3, \quad (3)$$

1st Order Median Moment

$$M\mu(I) = \begin{cases} SI\left(\frac{N^2}{2}\right), & N \text{ is odd} \\ \frac{SI\left(\frac{N^2-1}{2}\right) + SI\left(\frac{N^2+1}{2}\right)}{2}, & N \text{ is even} \end{cases} \quad (4)$$

Where SI is the sorted list of the pixel values of the 3×3 block

2nd Order Median Moment

$$Me_2(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (I(i, j) - M\mu(I))^2 \quad (5)$$

3rd Order Median Moment

$$Me_3(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (I(i, j) - M\mu(I))^3, \quad (6)$$

The average Mean and Median moments are calculated as

$$AMM_{0 < i < m} = \frac{M(i, j) + \mu_2(i, j) + \mu_3(i, j) + M_\mu(i, j) + Me_2(i, j) + Me_3(i, j)}{6}$$

(7)

Where m and n are the number of rows and columns in the image, respectively and N is 3 in the above equations.

C. Word Extraction

The Max-Min clustering gives a text cluster but that is still not sufficient to identify the gap between words due to the sparsity of the matrix. To restore the lost text information, we propose a union operation of the Canny edge map of the input image with the text cluster obtained by Max-Min clustering (Figure 2(e)) as it is known that the Canny operator definitely gives edges for text and that can be used for word segmentation (Figure 2(f)) but not for character segmentation due to erratic edges at the character background. Hence the union operation helps in filling the text region and leaving a gap between words as shown in Figure 2(g).

The gap between words is identified by introducing the concept of run length which is a well known method for segmentation of text in document analysis. This idea works when the image has a high number of identical consecutive black pixels (background). It is true that where there is a gap between words there will get consecutive black pixels in a high number but not in between characters. Hence, this idea gives good results for word extraction as shown in the sample results in Figure 2(h) where we observe a clear space between the words and correct segmentation of words.



(a). Input (b) Text candidates (c) Output
Figure 4. Character Extraction

D. Character Extraction

The run length concept used for word segmentation does not work for character extraction as it is noticed in Figure 4(b) which is the result of Max-Min clustering on Figure 4(a) that there is no such a high number of consecutive black pixels between characters. Therefore, we propose a new method based on text height difference (THd) vector, top distance and bottom distance vector of the above union operation. The THd is defined as the distance between the topmost pixel and the bottommost pixel of each column in the restored word image. If THd is less than two pixels, we consider the gap as a true character gap. If not then we check the top and bottom distance vectors. The top distance vector (Td) is defined as the distance between the upper boundary and the topmost pixel of each column and the bottom distance vector (Bd) is defined as the distance between the lower boundary and the bottommost pixel of each column. Then the method finds the difference between consecutive distance values in Td and Bd to identify the depth (high difference when character boundary exists between the characters), which is denoted as Dtd and Dbd respectively. When there is a gap between characters, both Dtd and Dbd give high values and hence it is considered as a candidate gap. It can be seen in Figure 5(a) and (b) that the candidate gaps are marked by red color dots for both Dtd and Dbd vectors, respectively. We have observed while doing experimentation that touching between two characters exists at the middle of character boundary but not at top and at bottom. For Dtd and Dbd values, we use the same Max-Min clustering method used in Section B for obtaining text cluster (Figure 4(b)) for choosing candidate gap clusters (Cluster with Max value). For each candidate gap in the cluster belonging to the top distance vector, the proposed method checks whether the corresponding candidate in the candidate gap cluster obtained from the bottom distance vector is also a candidate

gap or not. If so, we consider it a true candidate gap for extracting the character as shown in sample results in Figure 4(c) where the characters are segmented correctly.

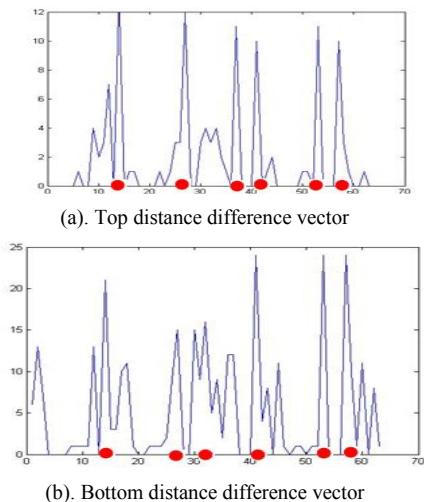


Figure 5. Candidates for character gap identification

The red ● dots indicate that those columns are candidates for character gaps.

III. EXPERIMENTAL RESULTS

As there is no standard database available for video text detection, we create our own dataset for the purpose of experimentation. We have selected 1216 text line including Chinese, 3310 words and 3527 characters. Our database includes different varieties of multi-oriented scene text line images in order to show that the proposed method is effective and is useful for video text recognition.

To evaluate performance of the proposed method, we consider recall (R), precision (P) and f-measure (F) as measures in this work. Experiments in terms of recall and precision for word and character extraction and recognition results in terms of recognition accuracy are presented in this work. We conduct experiments on the text cluster (Cluster) obtained by Max-Min clustering algorithm, Canny operation of the original text line image (Canny) and the restored image of the proposed method for word and character extraction (Union). Similarly, we conduct the same experiments on the Chinese data and the results are reported in Tables 2 and 4. The sample results for word and character extraction given by the proposed method is shown in Tables 1 and 3 where gaps are shown in white color for different varieties of word and character images. It is noticed from the results in Tables 1 and 3 that the proposed method works well for word and character extraction in scene video.

Table 2 shows that recall, precision and f-measure of the proposed method for word segmentation are higher than the results given by the text clustering algorithm and Canny operation alone. This is because text clustering loses significant text information when Max-Min clustering is used while Canny operation gives erratic edges due to the background complexity. On the other hand, the method with the union operation gives better results because of the advantage of segmentation of words and characters.

Table 4 shows character extraction results of the proposed method for different classes of data and it does not include results on text clustering and Canny operation as they have already been

shown to give poorer result at the word level and hence poorer result at the character level. It is noticed from Table 4 that the proposed method give slightly better results for horizontal text than non-horizontal text because the conversion algorithm from non-horizontal to horizontal does not always preserve the quality of the image.

Table 1. Sample results for word extraction

Gray Image	Result

Table 2. Word extraction for English and Chinese horizontal and non-horizontal data

Method	English Horizontal and non-horizontal			Chinese Horizontal and non-horizontal		
	R	P	F	R	P	F
Cluster	0.80	0.72	0.75	0.82	0.75	0.78
Canny	0.77	0.81	0.78	0.76	0.79	0.77
Cluster +Canny (Union)	0.85	0.87	0.85	0.86	0.88	0.86

Table 3. Sample results for character extraction

Gray Image	Result

Table 4. Character extraction for both English and Chinese data

Cluster + Canny (Union)	R	P	F
English Horizontal	0.85	0.87	0.86
English Non-Horizontal	0.81	0.84	0.82
Chinese Horizontal	0.87	0.88	0.88
Chinese Non-Horizontal	0.82	0.85	0.83

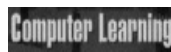







A. Recognition Results

We present recognition results on English data but not on Chinese data due to the non-availability of a Chinese OCR engine. Besides, our primary goal of this work is to show that extraction of character helps in improving accuracy in recognition. We use the latest method reported in [18] for binarization before passing to OCR. The reason behind in making the choice of this binarization

method is that this work has been compared with state of the art methods to prove its superiority to existing methods on binarization. The OCR engine found in [19] is used for the purpose of English text recognition.

To show that extraction of words and characters is useful in improving video text recognition with the current OCR, we present the recognition results at three levels namely text line (we feed the whole text line into OCR), word (we feed the extracted words into OCR) and character (we feed the extracted individual characters into OCR) and the results are reported respectively in the first, second, and third and fourth rows in Tables 5. Table 5 shows that the current OCR performs poorly at the text line level as a low character recognition accuracy of 36.5% is reported (First row). A sample recognition result at the word level is given in the second row in Table 5 where one can notice less errors in recognition results and that a higher character recognition accuracy of 62.4% is reported compared to the recognition accuracy at the text line level due to elimination of complex background by the segmentation method. As a result, it can be concluded that word extraction helps in improving the recognition accuracy. Table 5, the third and fourth rows show sample results of two character images which have been correctly recognized. The recognition accuracy at the character level is 65.6% which is higher than that at the text line and word levels. We also observe that the recognition accuracy is not much different at the word and character levels due to the language model in use at the word level that helps resolving ambiguity in recognizing the characters.

Table 5 .Recognition results at text line, word and character level

Input	Binarization	Recognition
		Nrrmum Laaning
		PALLUJHA
		P
		F

IV. CONCLUSION AND FUTURE WORK

In this work, we have proposed a novel segmentation method based on Fourier-moments features for word and character extraction from text line image and word image. We have shown that the combination of Fourier, mean and moments are good for classification of text pixels. The run length concept is applied for the first time on word gap identification in video from the restored image. Novel distance vectors are proposed for character extraction from words. The experimental results of the recognition reveal that extraction of words and character is useful to improve the accuracy of OCR recognition. In future, to achieve better accuracy as in document analysis, we are planning to develop a reconstruction

algorithm to restore character shapes for extracted characters from text line images.

ACKNOWLEDGMENT

This work is supported in part by A*STAR grant R252-000-402-305.

REFERENCES

- [1] S. H. Lee and J. H. Kim, "Complementary combination of holistic and component analysis for recognition of low resolution video character images", *Pattern Recognition Letters*, 2008, pp 383-391.
- [2] D. Chen and J. M. Odobez, "Video text recognition using sequential Monte Carlo and error voting methods", *Pattern Recognition Letters*, 2005, pp 1386-1403.
- [3] D. Chen, J. M. Odobez and H. Bourland, "Text detection and Recognition in images and video frames", *Pattern Recognition*, 2004, pp 595-608.
- [4] X. Tang, X. Gao, J. Liu and H. Zhang, "A Spatial-Temporal Approach for Video Caption Detection and Recognition", *IEEE Transactions on Neural Networks*, 2002, pp 961-971.
- [5] D. Doermann, J. Liang and H. Li, "Progress in Camera-Based Document Image Analysis", In *Proc. ICDAR 2003*, pp 606-616.
- [6] C. Wolf and J. M Jolion, "Extraction and Recognition of artificial text in multimedia documents", *Pattern Analysis and Applications*, 2003, pp 309-326.
- [7] X. Hunag, H. Ma and H. Zhang, "A New Video Text Extraction Approach", In *Proc. ICME 2009*, pp 650-653.
- [8] J. Zang and R. Kasturi, "Extraction of Text Objects in Video Documents: Recent Progress", In *Proc. DAS 2008*, pp 5-17.
- [9] K. Jung, K.I. Kim and A.K. Jain, "Text information extraction in images and video: a survey", *Pattern Recognition*, 2004, pp. 977-997.
- [10] A.K. Jain and B. Yu, "Automatic Text Location in Images and Video Frames", *Pattern Recognition*, 1998, pp. 2055-2076.
- [11] H. Li, D. Doermann and O. Kia, "Automatic Text Detection and Tracking in Digital Video", *IEEE Transactions on Image Processing*, 2000, pp 147-156.
- [12] K. L. Kim, K. Jung and J. H. Kim, "Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm", *IEEE Transactions on PAMI*, 2003, pp 1631-1639.
- [13] P. Shivakumara, T. Q. Phan and C. L. Tan, "A Laplacian Approach to Multi-Oriented Text Detection in Video", *IEEE Transactions on PAMI*, 2011, pp 412-419.
- [14] Z. Saidane and C. Garcia, "Robust Binarization for Video Text Recognition", In *Proc. ICDAR 2007*, pp 874-879.
- [15] Z. Zhou, L. Li and C. L. Tan, "Edge based Binarization for Video Text Images", In *Proc. ICPR 2010*, pp 133-136.
- [16] K. Jung, "Neural network-based text location in color images", *Pattern Recognition Letters*, 2001, pp 1503-1515.
- [17] Donald Hearn and M. Pauline Baker, "Computer Graphics C Version" 2nd Edition, Prentice-Hall, 1994, Bresenham Line Drawing Algorithm.
- [18] S. Bolan, L. Shijian and Chew Lim Tan. "Binarization of Historical Document Images Using the Local Maximum and Minimum". In *Proc. DAS 2010*, pp 159-165.
- [19] OCR Engine used: <http://code.google.com/p/tesseract-ocr/>