Hybrid Chinese/English Text Detection in Images and Video Frames

Wenge Mao, Fu-lai Chung[†]
Centre for Multimedia Signal Processing and
Dept. of Computing
Hong Kong Polytechnic University
Hung Hom, Hong Kong.

Kenneth K.M. Lam and Wan-chi Siu
Centre for Multimedia Signal Processing and
Dept. of Electronic and Information Engineering
Hong Kong Polytechnic University
Hung Hom, Hong Kong.

Abstract

In this paper, we propose a multiscale texture-based method using local energy analysis for hybrid Chinese/English text detection in images and video frames. Local energy analysis has been shown to work well in text detection, where remarkable local energy variations of pixels correspond to text region or boundary of other objects and lower local energy variations of pixels correspond to background or the interior of non-text objects. Local energy variation is calculated in a local region based on the wavelet transform coefficients of images. Hybrid Chinese/English text in images and video frames can be detected whether it is aligned horizontally or vertically. The font size of text to be detected may vary in a wide range of values. The proposed method has been tested on 321 frame images obtained from local TV programs and a tested dataset with low missed rate and false alarm rate.

1. Introduction

Text embedded in video frames provides a powerful cue for video database indexing and retrieval. However, due to the complexity of its appearance and the background in video frames, text detection is still a difficult and yet challenging task in computer vision. Some research efforts have been done in the past several years and most of the proposed approaches are based on unsupervised classification, such as [3,7,8,9], in which the text is characterized by texture analysis and/or connected component analysis, while others are based on supervised learning, such as [4,5].

In this paper, we mainly focus on hybrid Chinese/English text detection and propose a method based on local energy analysis of pixels in still images and video frames. Due to the difficulty of achieving optimal selection of parameters in a bank of Gabor filters [1] in the case of a wide range of font size, we use 2D Harr wavelet transform to characterize the local energy variations of pixels instead of a bank of Gabor filters. The pixels within text-like objects and near the boundary of other objects have large local energy variations, but the pixels within the background and the interior of non text-like objects have relatively smaller local energy variations. This is the criterion of discriminating between text regions, other objects and the background. The reminder of this paper is organized as follows. Section 2 presents the text detection method. Section 3 discusses the experiment results and the conclusion is provided in the final section.

2. Text detection using wavelet transform and local energy variation

Prior to describing the details, our text detection method is outlined here with the help of the system architecture depicted in Fig.1. In the first step, wavelet transform of an image is done to characterize the local energy variations (LEV) of pixels in the successive scale levels. In each scale level, the corresponding local energy variations are computed and then are thresholded. The resulting binary map image in each scale level is subsequently analyzed by connected component analysis (CCA) technique to label different objects and background. Since some characters touch other objects, such characters can be separated from the touching object using the projection profile of an isolated connected component. Further, text regions are located from other connected object regions by the predefined geometric filtering. In the final step, all text regions in the consecutive scale levels are fused into the original image and text regions are detected. Note that these processing steps are performed on grayscale images, not on color images.

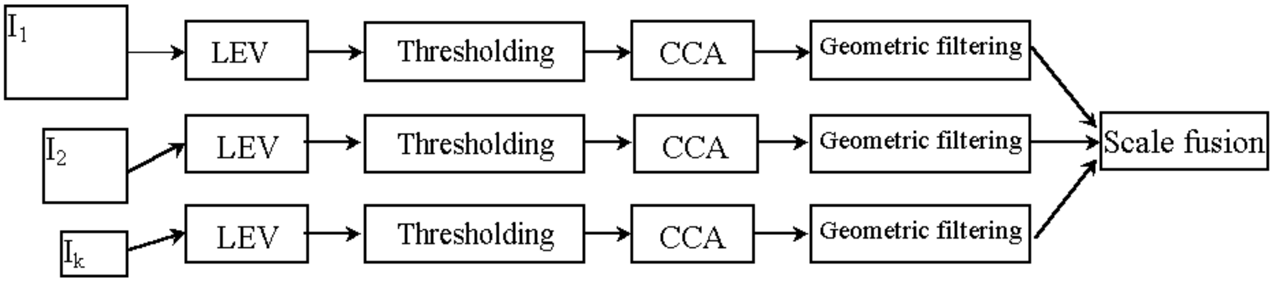


Fig.1 System architecture for text detection

[†] *Corresponding author (cskchung@comp.polyu.edu.hk)*

1051-4651/02 $17.00 (c) 2002 IEEE

2.1. Local energy variations (LEV)

It is vital to text detection how to characterize the texture feature of text and further to discriminate between text regions, other objects and background. Wavelet transform is more powerful to do this than conventional differential filters. In [6], the selection of an optimal wavelet basis for texture characterization is discussed. Experiments have shown that Harr wavelet basis possesses rather good ability to characterize texture features and is computationally efficient. More importantly, Harr wavelet decomposition can avoid or lessen boundary effect that may result in spurious boundary in the image borders. We here adopt tensor Harr wavelet basis to decompose images into the successive scale levels. As shown in Fig.1, I_1, I_2 and I_k are the transformed coefficients in the 1st, 2nd, and *k*-th scale levels, respectively.

Let $\varphi(x)$ and $\psi(x)$ be the Harr scaling and wavelet functions of variable *x*, respectively, then the functions by dilations and translations of them are:

$$\varphi_{k,i}(x) = 2^{-k/2}\varphi(2^{-k}x - i)$$
$$\psi_{k,i}(x) = 2^{-k/2}\psi(2^{-k}x - i) \qquad (1)$$

In the 2D case, the corresponding transform basis functions in tensor product form are as follows:

$$\varphi^{LL}_{k;i,j}(x,y) = \varphi_{k,i}(x)\varphi_{k,j}(y)$$
$$\psi^{LH}_{k;i,j}(x,y) = \psi_{k,i}(x)\varphi_{k,j}(y)$$
$$\psi^{HL}_{k;i,j}(x,y) = \varphi_{k,i}(x)\psi_{k,j}(y)$$
$$\psi^{HH}_{k;i,j}(x,y) = \psi_{k,i}(x)\psi_{k,j}(y) \qquad (2)$$

Let $f(x,y)$ be the intensity of a pixel at (x,y), where *x* varies from 0 to $M_0 - 1$, *y* from 0 to $N_0 - 1$, M_0 and N_0 are the height and width of images respectively. In each scale level, the image is decomposed into four subbands: *LL*, *LH*, *HL*, and *HH*, where *LL* represents the horizontally low-frequent and vertically low-frequent components of the image, *LH* the horizontally low-frequent and vertically high-frequent components, *HL* the horizontally high-frequent and vertically low-frequent components and *HH* the horizontally high-frequent and vertically high-frequent components. The three high-frequent subbands can be used to characterize the intensity variation of neighboring pixels in the image. We here adopt *LH* and *HL* coefficients to extract local texture features of the image. Let $LH^k(x,y)$ and $HL^k(x,y)$ be the Harr wavelet transform coefficients in the *k*-th scale level, where $k = 1,\cdots,L$; $x = 0,\cdots,M_k - 1$ and $M_k = \lfloor (M_{k-1}+1)/2 \rfloor$; $y = 0,\cdots,N_k - 1$ and $N_k = \lfloor (N_{k-1}+1)/2 \rfloor$, $\lfloor N \rfloor$ denotes the largest integer smaller than *N*. Let $LIV^k(x,y)$ be the local intensity variation of (x,y) in the *kth* scale level and $LEV^k(x,y)$ the local energy variation of (x,y) in the same scale level. Then, it follows that

$$LIV^k(x,y) = abs(LH^k(x,y)) + abs(HL^k(x,y)) \qquad (3)$$

and

$$LEV^k(x,y) = \max_{(m,n)\in D}(LIV^k(m,n)) - \min_{(m,n)\in D}(LIV^k(m,n)) \qquad (4)$$

where D is a small region with fixed size containing (x,y) and is often selected to be a rectangular region.

From eq.(3) and eq.(4), it can be seen that the boundary pixels of objects will have large local energy variations while the pixels in the background or far away from the object boundaries will have small local energy variations. The size of region D should be set appropriately. If it is too large, more non-boundary pixels will have large local energy variations and it will be difficult to differentiate character regions from the boundary regions. If the size of D is too small, the character components will break after thresholding and character regions may be in a clutter.

The LEV analyzed images consist of three parts: (i) regions in which pixels have small local energy variations, corresponding to the background or the interior far away from the boundary of non text-like objects; (ii) regions in which pixels have large local energy variations, corresponding to the boundary of non-text objects; and (iii) regions in which pixels have large local energy variations, corresponding to text-like objects. The boundary with large local energy variation is usually slim and can easily be filtered out in the later steps. However, character-like objects are grouped into clusters with strip shape or approximately square shape and these are being exploited in the later detection step.

2.2. Thresholding

After the above step, the original image is completely characterized by its LEV analyzed images. The local energy variation for boundary of objects and characters is larger than that of the background and the interior of non-text objects. A thresholding step is thus adopted here to discriminate between them. Characters with high contrast have very large local energy variations and are easily perceivable by viewers. On the other hand, characters with low contrast do not have very large local energy variations and have to be recognized attentively. In those few cases, characters are unrecognizable since their contrasts are too low and their local energy variations are almost the same as that of background. Based on such observations, the local energy variations of pixels within characters may cover a large range of values and an appropriate threshold value should not be very high. We here select an adaptive threshold value to perform global thresholding for text detection. The threshold value is set at a certain percentage of the largest local energy variation in an image. The ratio of the threshold value to the largest local energy variation was chosen to be 0.45 in our experiments and it is effective in separating characters with not too low contrast from the other parts of the images.

2.3. Connected component analysis (CCA) and geometric filtering

The result of thresholding an LEV analyzed image in each scale level is a binary map image, in which pixels with value 1 correspond to large local energy variations and pixels with value 0 denote low local energy variations. Pixels with value 1 form the foreground of the image, which can be further divided into three parts, namely, text-like regions, slim boundary regions and small noise regions. There will exist many different connected components and a connected component analysis is necessary to label the components for subsequent analysis. Due to the potential touching between characters and other objects, the connected components containing characters may be enlarged by the touching objects and need be refined. Either the horizontal or vertical projection profile is computed according to the size of the minimum bounding rectangle (MBR) of the components. If the width of the MBR is larger than the height at a predefined rate, the horizontal projection is performed and the connected components are segmented into separate small parts when the projection profile satisfy some conditions (omitted here due to space limit). Otherwise, the vertical projection is done in the same process as the horizontal projection.

Geometric filtering follows the CCA process. Texts are subjected to certain geometric restrictions. Their height, width, aspect ratio and compactness do not take any fixed value but generally fall into a wide range of values. In this paper, we mainly consider two cases, namely, horizontal text lines and vertical text lines. The geometric restrictions of a potential text line are as follows:

- The height of a horizontal text line must be greater than 6 pixels in the map image; and
- The width of a vertical text line must be greater than 6 pixels in the map image; and
- The filled ratio of a text line in its MBR must be over 0.5.

If a region satisfies all these three requirements, it is inferred to be a potential text region.

2.4. Multi-scale fusion

In this step, the potential text regions detected in the consecutive scale levels can be integrated into the original image. In each scale level, the local energy variation image can be restricted only to process a fixed range of heights of characters. The characters outside the range are discarded and thus the filtering in different scale levels corresponds to a channel that detects characters of their heights within the stipulated range.

3. Experimental results

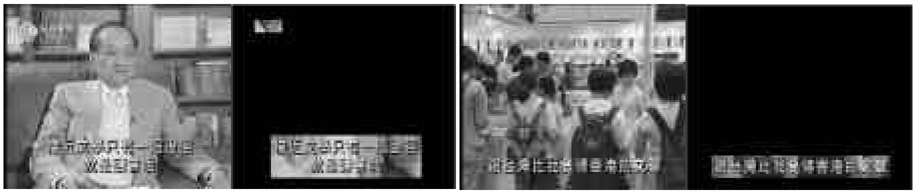
To evaluate the performance of the proposed method, we have tested it on a number of images, including still images and video frames captured from Hong Kong local TV programs. The video frames are mainly about news, literature, dancing, visual art and Chinese traditional opera. We have tested 321 images, in which Chinese characters mainly appear in 271 of them and English characters mainly appear in the rest [2]. The number of Chinese characters within them is 3533 and the number of English characters is 1361. The font sizes of characters in images vary from 12pt to 64pt. Most of characters are aligned horizontally but there are some aligned vertically. Some characters are laid over simple background. However, there is still a lot of characters laid over complex background. By experiments, we found that the contrast of characters with their local background extends a wide range in some images.

Table 1: The result of text detection

Text	Chinese	English
Number of characters	3533	1361
Correctly detected	3147	1317
Missed	386	44
Number of false alarm regions	570	
Missed rate (%)	10.9	3.2

As shown in Table 1, the proposed method achieved missed rates of 10.9% among all the Chinese characters and 3.2% among all the English characters in the tested images. The major causes for the false rejects are: i) the contrast of the false rejected characters is too low to be detected because the artificial characters are laid over the complex background with significantly perturbed intensities; and ii) the complicated connection with the other objects so that simple separation of characters from other objects fails and they are falsely discarded. The false alarm of regions is due to the fact that the false alarm regions have strong texture features and high contrast and resemble text-like regions.

Fig.2 demonstrates the case that text lines are in horizontal orientation. Here, Fig.2 (a) depicts the results of text detection with simple background while Fig.2 (b) shows the results of text detection with complex background. The results of vertical text lines are shown in Fig.3. Non-artificial text detection and hybrid Chinese/English text detection are shown in Fig. 4.



(a) (b)

Fig. 2. The results of artificial text detection (horizontal orientation): (a) with simple background; (b) with complex background.

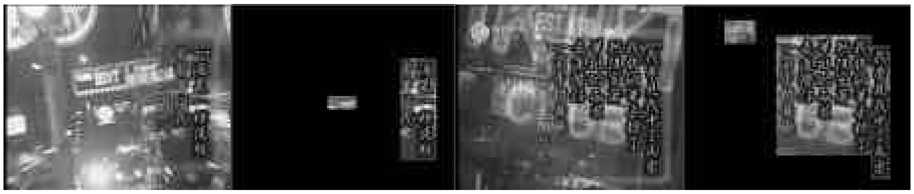


Fig.3. The results of artificial text detection (vertical orientation)

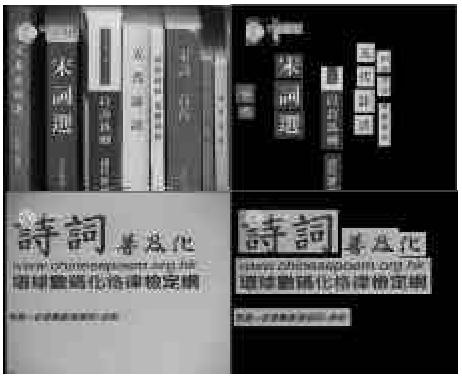


Fig. 4. The results of scene text detection and hybrid Chinese/English text detection

4. Conclusions

In this paper, we have proposed a method to detect hybrid Chinese/English text in images and video frames. The method is based on the wavelet transform and local energy variation analysis to discriminate between text-like regions, boundary and interior of other objects, and the background. It can effectively detect text regions within images whether they are aligned horizontally or vertically. Furthermore, the method can simultaneously detect characters of various font sizes in a single image. The method works well for the tested dataset with low missed rate and low false alarm rate. The idea of local energy variation analysis for text detection deserves to be further studied.

Acknowledgement

The authors would like to acknowledge the support of this project by the Multimedia Signal Processing Centre of Hong Kong Polytechnic University under grant A430.

References

[1] W. Chan and G. Coghill, “Text analysis using local energy,” *Pattern Recognition*, vol. 34, pp. 2523-2532, 2001.

[2] X. Hua, W. Liu and H. Zhang, “Automatic performance evaluation for video text detection,” *Proc. of the Sixth Int. Conf. on Document Analysis and Recognition (ICDAR)*, pp.545-550, Sept. 10-13, 2001.

[3] A.K. Jain and B. Yu, “Automatic text location in images and video frames,” *Pattern Recognition*, vol.31, no.12, pp.2055-2076, 1998.

[4] K.I. Kim, K. Jung, S.H. Park, and H.J. Kim, “Support vector machine-based text detection in digital video,” *Pattern Recognition*, vol.34, no.2, pp.527-529, 2001.

[5] H. Li, D. Doermann, and O. Kia, “Automatic text detection and tracking in digital video,” *IEEE Trans. on Image Processing*, vol.9, no.1, pp.147-156, 2000.

[6] A. Mojsiloviæ, M.V. Popoviæ and D.M. Rackov, “On the Selection of an Optimal Wavelet Basis for Texture Characterization,” *IEEE Trans. on Image Processing*, vol.9, no.12, pp.2043-2050, Dec. 2000.

[7] T. Sato, T. Kanade, E. Hughes, M. Smith, “Video OCR for digital news archives,” *IEEE Workshop on Content-Based Access of Image and Video Databases (CAIVD’98)*, Bombay, India, January 1998.

[8] V. Wu, R. Manmatha, E.M. Riseman, “TextFinder: An automatic system to detect and recognize text in images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.21, no.11, pp.1224-1229, 1999.

[9] Y. Zhong, H.J. Zhang, and A.K. Jain, “Automatic caption localization in compressed video,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.22, no.4, pp.385-392, 2000.