http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (8): 1527-1535, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1527.1535

Robust Text Binarization Based on Line-Traversing for News Video

[1]Shwu-Huey Yen, [1,2]Hsiao-Wei Chang, [1]Chia-Jen Wang and [1]Chun-Wei Wang
[1]Department of Computer Science and Information Engineering, Tamkang University,
151 Ying-Chuan Road, Tamsui, Taipei County 25137, Taiwan, Republic of China
[2]Department of Computer Science and Information Engineering,
China University of Science and Technology, 245, Sec. 3, Academia Road, Taipei City 11581,
Taiwan, Republic of China

Abstract: This study presents a robust approach to binarize the detected rectangular text regions (text boxes) on news videos. The binarization problem can be traced back to 1970s but it is still challenging in news video today since the background is complicated and unpredictable. The proposed algorithm adopts the line traversing method integrating the edge information and the intensity statistics to accomplish the binarization task. First, Canny edge detector is applied on a text box. Next, the vertical line scanning from left to right of the text box is performed twice. The vertical line traverses downwards until it hits an edge pixel or reaches the bottom of the box. Similarly, the vertical line traverses upwards until it hits an edge pixel or reaches the top of the box. These traversed pixels are classified as background pixels. From the histogram of those non-background pixels, the peak intensity p and the standard deviation σ are evaluated. The threshold for text in news video is set to be T = (0, p+kσ) or T = (p-kσ, 255), depending on text polarity. In the case that the range of background intensity covers the entire intensity range of the image, the algorithm uses the temporal information of news video to remove most of the background. Moreover, the intensities of those background pixels, whose intensity is similar to the text pixels, are replaced by 255 or 0, depending on the text polarity. Finally, a binarization is performed in this modified text box. Notice that the proposed method is parameter-free, has no limitation on the text polarity and can handle the case of similar intensity in background and text for news video. The method has been extensively experimented on text boxes from various news videos, historical archive documents and other different documents. The proposed algorithm outperforms the well-known methods such as Otsu, Niblack, Sauvola, etc., in speed, precision and quality.

Key words: Text extraction, binarization, canny edge, text polarity, otsu, niblack

INTRODUCTION

Digital video plays an important role in our life these days, encompassing entertainment, education and other multimedia applications. With massive video files in hand, an effective tool for users to browse, index and obtain video contents is crucial. Among the various video analysis techniques proposed, text is one of the most common video mediums. Not only is text a well-known and accustomed tool to people but also text contains a wealth of high-level semantic information that can be used to conduct video retrieval or video annotation. Therefore, the study of video textual information extraction is a valuable work. Text in news video is especially significant to the semantic content of the news. In general, text may appear as scene text or caption text (Jung *et al*., 2004). An example of scene text could be a store's name in a news report on rising prices in daily necessities. The name of this store is not related to the content. In fact, most scene texts are not that important of a factor when it comes to video content. Alternately, caption text, static or scrolling, is superimposed in the later stage of video production. Most static texts provide concise and direct descriptions of the content presented in news videos such as the title of the current issue, whereas scrolling texts are usually updated information such as stock markets figures and upcoming programs which are not related to the current content. The focus of this study is on static text on the news videos.

The procedure of video textual information extraction can be broadly divided into two categories: detection and extraction. Detection is to label (locate) text regions and extraction is to binarize the detected text regions and to perform Optical Character Recognition (OCR). Text detection in a video sequence is challenging due to complex background. Much research effort has been

Corresponding Author: Hsiao-Wei Chang, Department of Computer Science and Information Engineering, China University of Science and Technology, 245, Sec. 3, Academia Road, Taipei City 11581, Taiwan, Republic of China

Fig. 1: The detected news video text blocks of the algorithm by Yen *et al.* (2010)

dedicated to this issue such as Wang *et al.* (2004), Lyu *et al.* (2005) and Yen *et al.* (2010). Only good text detection can lead to successful text extraction. In order to determine the quality of a resulting detected text box in a video sequence, the following requirements must be met: the text is masked or bounded by a box as tight/close as possible, the false detection is low and the recall rate is high. Based on the integration of multiple frames, an algorithm to fulfill these requirements was proposed in the previous work (Yen *et al*., 2010). One of the novelties in Yen *et al.* (2010) is that most of the irrelevant background is eliminated in the resulting text box. In the previous proposed algorithm, four reference frames are taken in every 30 frames to perform Canny edge detection. Operations of logical AND on resulting Canny edge maps and line-deletion in rough detected boxes contribute to the clean text preserved in the refined detected boxes. Figure 1 gives two examples using the algorithm in Yen *et al.* (2010); Fig. 1b and f are refined results where, Fig. 1a and e are the original color video frames. Comparing text boxes in Fig. 1d and h (refined ones) to Fig. 1c and g (original gray images), it can be seen that most of the background has been cleared. Notice that in Yen *et al.* (2010), the polarity of the refined result is reversed since OCR processes text in positive polarity (dark text on bright background) and most video text is negative polarity (bright text on dark background). More details can be found in Yen *et al.* (2010). The purpose of this study is to binarize detected text boxes; it is a continuation of the previous work. To justify the generality of the proposed binarization scheme, we tested the algorithm on various detected text boxes as well as historical archive documents.

RELATED WORK

The study of text binarization methods (Ntogas and Ventzas, 2008) has a long history and it can be grouped into two broad categories: global binarization (a single threshold is calculated for the whole image) and local binarization (local thresholds are calculated within separate windows or areas). Two global approaches proposed by Otsu (1979) and Lin and Yang (2000) are briefly reviewed. Otsu (1979) detects text from grayscale

image using clustering analysis which creates two clusters of Gaussian distribution of the pixels as background and foreground regions. The optimal threshold value is obtained by minimizing the weighted sum of within-class variance of the two classes of pixels. Lin and Yang (2000) proposed an intuitive histogram-based method of image thresholding. The cluster peaks of a histogram are determined based on mountain clustering method (Yager and Filev, 1994) which estimates cluster centers on the basis of a density measure. When text has similar intensity to the background's or is printed on a background of non-uniform brightness, global binarization is prone to have unsatisfactory results. Local approaches can make up for these problems in general.

Niblack (1986) is one of the most well-known local approaches. In evaluating eleven different locally adaptive binarization methods for gray scale images with low contrast, variable background intensity and noise, Trier and Jain (1995) concluded that Niblack's method performs better than other local thresholding methods. To determine whether a pixel is a text in Niblack's method, the threshold is decided utilizing local mean and standard deviation for the window centered at this pixel as in Eq. 1:

$$T(x, y) = m(x, y) + k * s(x, y) \quad (1)$$

where, m (x, y) and s (x, y) are local mean and standard deviation values of the window, respectively. The size of the window is determined empirically and the value k is a constant which determines how much of the total edge pixels are retained. As commented in Wolf *et al.* (2002) and Gatos *et al.* (2006) and other papers, k is suggested to be -0.2 and the results are not very sensitive to the window size as long as the window covers at least 1-2 characters. However, in addition to the complicated background, various character sizes appearing in one frame is very common to news video. Present experiments indicate that choosing suitable parameters in Niblack algorithm for video text is not an obvious task. Besides, Niblack method has a drawback-noise created in areas that do not contain any text (due to the fact that a threshold is created in these cases as well). Sauvola *et al.* (1997) solved this problem by adding a hypothesis on the gray values of text (near 0) and background pixels (near 255) which results in Eq. 2 for the threshold:

$$T(x, y) = m(x, y)*(1+k*(s(x, y)/R-1)) \quad (2)$$

where, m (x, y) and s (x, y) are as in Niblack's formula. R is the dynamic range of standard deviation fixed to 128 and k is a parameter. Sezgin and Sankur (2004) have evaluated 40 different thresholding methods in the applications of non-destructive testing and document analysis. They conclude that Sauvola's binarization method works better than other local binarization techniques. Nevertheless, it has similar problem as in Niblack's that it is not easy to choose suitable parameters k and window size in video text binarization. Since intensities of text and background may not correspond to the hypothesis when binarization takes place Wolf *et al.* (2002) improve Sauvola's formula in order to normalize the contrast and the mean gray level of the image. The threshold formula is given in Eq. 3:

$$T = m-k\alpha (m-M), \alpha = 1- s/R \quad (3)$$

where, M is the minimum graylevel of the image and R = max (s), the maximum of the standard deviations of all windows. Based on Niblack's and Sauvola's algorithms, He *et al.* (2005) modifies both algorithms in binarization of historical archive documents which are poor in quality due to age, discolored cards and ink fading. They propose an adaptive Niblack's algorithm which allows choosing values for parameter k and window size automatically. They also gave an adaptive version of Sauvola's algorithm where the window size is half of the height of the input characters and that the k value should be chosen adaptively for each small window. Ngo and Chan (2005) proposed another adaptive thresholding approach that claims being capable of reducing most noise caused by complex background scenes. In their approach, a 16-bin normalized histogram of the grey level values is computed first. Then, by scanning the bins backward, bin k that corresponds to the first valley in front of the first peak is located. Finally, a threshold with a value of $16\times(k-1)$ is used to binarize the text image.

THE PROPOSED METHOD

The study proposes a robust and simple binarization algorithm for detected text blocks of news videos. As mentioned in Section one, we can obtain a tight detected text boxes with most of background removed (Fig. 1d, h) from the previous work (Yen *et al.*, 2010). However, to make the algorithm an independent one, the algorithm presented here is based on text boxes in general such as in Fig. 1a and e. The flowchart of the proposed approach is displayed in Fig. 2.

Since a detected text block may tightly connect to text pixels as in Fig. 3a, we extend both the upper and lower boundary of the box 1 pixel as in Fig. 3b to ensure that those top/bottom margin pixels of the detected box are background pixels. Then, we apply Eq. 4 to transform the extended color image into grayscale image as in Fig. 3c:

$$Y = 0.299R + 0.587G + 0.114B \quad (4)$$

where, Y is the intensity value and R, G and B are the values on red, green and blue channels of the pixel.Based on this gray image, an intensity histogram and a Canny edge map are produced as shown in Fig. 4a and 3d.

Next, using the gray image and the corresponding Canny edge map, background pixels are roughly identified. For the gray image, a column-wise scanning from left to right is employed twice. Starting from the upper left position, the scanning is going downwards until it reaches a point which is also an edge point in the Canny edge map or the end of the boundary if there is no edge. Again, the scanning is repeated starting from the lower left position going upwards. Those encountered pixels are identified as background pixels. We let b_{min} and b_{max} be the least and the largest intensity values for those encountered pixels. The range of the background intensity is thus defined as B = (b_{min}, b_{max}). It is possible that B covers the entire range of the intensity histogram if background contains multiple colors including text colors. According to B, we will discuss the binarization in two cases.

Case 1: B does not cover the entire range of the histogram.

In our extensive experiments, this is the most common case. The interval B determines the range of foreground intensity F. Since text polarity is initially unknown, the midpoint of B can be used to estimate the foreground intensity. If intensity of the midpoint of B is bright (>128), then the text is positive polarity and negative polarity otherwise. The range of foreground intensity F is determined as in Eq. 5:

$$F = \begin{cases} [0, b_{min} - 1] & \frac{1}{2}(b_{min} + b_{max}) > 128, \\ [b_{max} + 1, 255] & \text{othewise}, \end{cases} \quad (5)$$

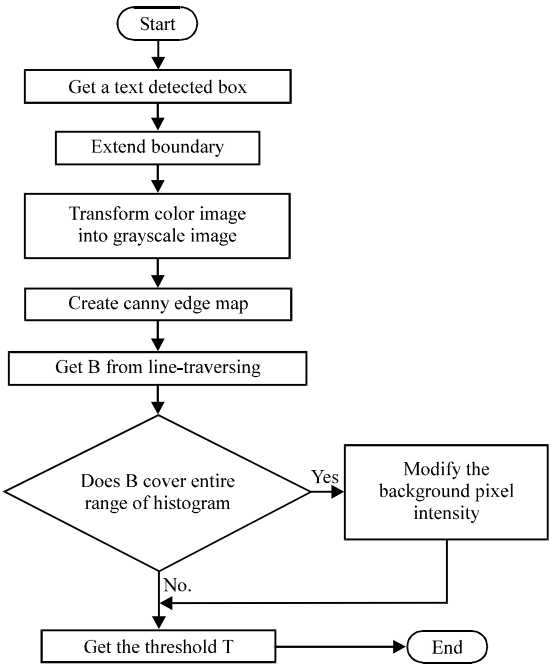

Fig. 2: The flowchart of the proposed approach

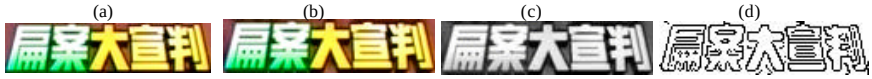

Fig. 3: (a) Original image, (b) extended image, (c) grayscale image and (d) Canny edge map

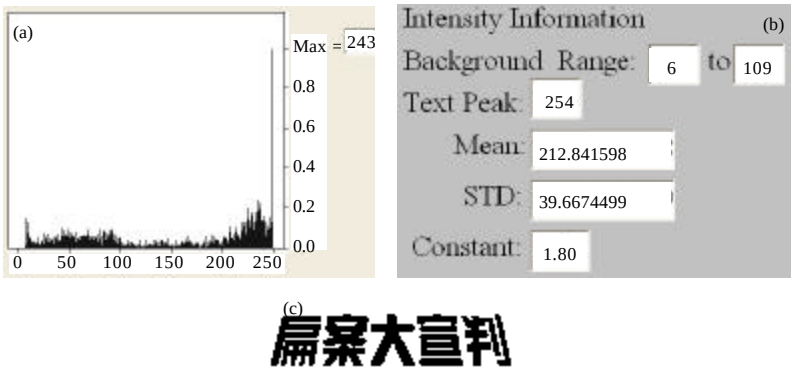

Fig. 4: (a) Histogram of the Fig. 3c, (b) intensity information of the Fig. 3c, (c) binarization result

where, b_{min} and b_{max} are from the background intensity range B. For those intensities within F, the peak p and the standard deviation σ are evaluated. As in Fig. 4b, it illustrates that the range of background intensity B is (6, 109) and the range of foreground intensity is therefore F = (110, 255). For intensities within F, the peak intensity p = 254 and the standard deviation σ = 39.67 are evaluated. Thus, we define the text intensity T as Eq. 6:

$$T=\begin{cases}[0, p+k\sigma] & \text{if it is positive polarity (dark text),}\\ [p-k\sigma,255] & \text{otherwisw,}\end{cases} \quad (6)$$

where, p and σ are the peak and the standard deviation of F and k is a constant. In our experiments, k is set as 1.80 and Fig. 4c shows the binarization result.

Case 2: B covers the entire range of the histogram.

Figure 5 exhibits an example. According to the gray image and its Canny edge map, the background intensity B = (0, 254) covers the entire range of the histogram. Most existing methods can not handle this kind of binarization problem. As depicted in Fig. 6, binarization results by methods of Otsu (1979), Lin and Yang (2000), Niblack (1986) and Sauvola *et al.* (1997) are not satisfactory.

To solve this problem, we use refined results from our previous work (Yen *et al.*, 2010) to obtain the threshold. Figure 7a shows the refined result of the detected text box corresponding to Fig. 5a, notice that it is in reversed text polarity as mentioned in section one. Since most of background are cleared (in white), the reversed image (Fig. 7b) shows a more homogeneous background comparing to that of Fig. 5a. We repeat the steps on this reversed image, i.e., producing Canny edge map, histogram and the intensity information. Figure 7c-e illustrate the results.

As in Fig. 7e, the background intensity range B is (0, 97) (comparing to (0, 254) in Fig. 5c) and it does not cover the entire range of histogram. Therefore, the text intensity is T = (174, 255) by Eq. 6. Although the refined image helps to provide a better background/foreground range estimation, it also causes artifacts near the conjunctions of text and background. Consequently, a fine tuning is implemented. First, taking Canny edge map of Fig. 5a as a reference, we again do column-wise scanning twice to label background pixels as before. For each background pixel P, if its intensity in Fig. 5a is within the range of text intensity T, then its intensity value will be replaced by Eq. 7:

$$I(P)=\begin{cases}255 & \text{if it is positive polarity,}\\ 0 & \text{otherwise,}\end{cases} \quad (7)$$

where, I (P) is the intensity of P. Figure 8a shows the tuned result. Comparing to Fig. 7b, the artifacts near the conjunctions of text and background are much reduced. The final binarization result is obtained by applying the threshold T to this modified image as shown in Fig. 8b. We also apply existing methods to the tuned image as displayed in Fig. 9. Overall, the binarization qualities are

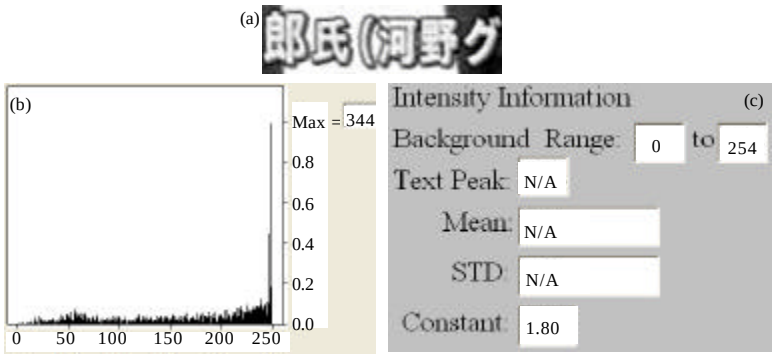

Fig. 5: (a) Original gray image (the same as in Fig. 1g), (b) and (c) are the corresponding histogram and intensity information

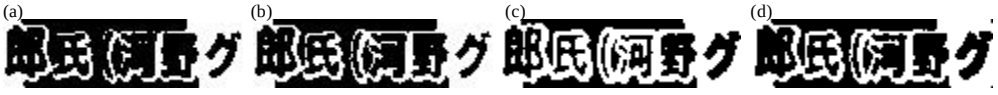

Fig. 6: The binarization results of Fig. 5a by (a) Otsu; (b) Lin; (c) Niblack (w = 30, k = 0.2); (d) Sauvola (w = 30, k = 0.2)

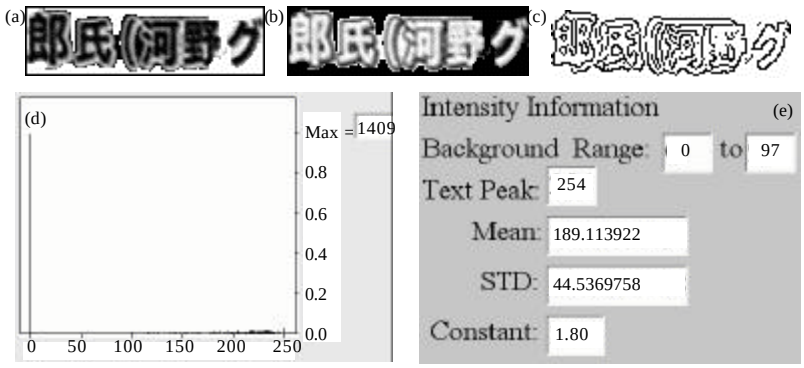

Fig. 7: (a) Detected text region (Fig. 1h), (b) reverse it; and (c) (d) (e) are Canny edge map, histogram and intensity information of (b)

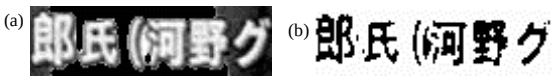

Fig. 8: (a) Modify background intensities of Fig. 5a and (b) proposed algorithm binarization result

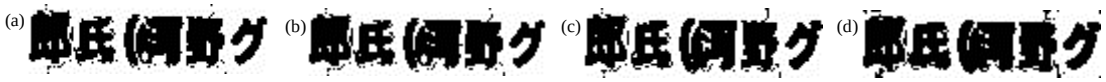

Fig. 9: The binarization result of Fig. 8a by method of (a) Otsu; (b) Lin; (c) Niblack (w = 30, k = 0.2) and (d) Sauvola (w = 30, k = 0.2)

improved comparing to the ones in Fig. 6. These binarization results confirm the effect of the modification on background intensity and our result (Fig. 8b) is the clearest one among them.

THE EXPERIMENTAL RESULTS

The following experiments illustrate that the proposed algorithm works on text boxes that are on news videos and on degraded documents. The comparisons are made between our algorithm to methods from Otsu (1979), Lin and Yang (2000), Niblack (1986), Sauvola *et al.* (1997), He *et al.* (2005) and Ngo and Chan (2005). Methods of ours, Otsu's and Lin's do not need to set any parameters. The parameters in Niblack's and Sauvola's are chosen from those who present the best binarization results by extensive trials and errors. The notations w and k indicate that the parameters used in the tests in which w = n means the window size is n×n. Due to the lack of original test images in He *et al.* (2005) and Ngo and Chan (2005), we can only show the figures given in the articles.

Figure 10 provides a summarization of some test results from the extensive tests done. Images in Fig. 10a-d are from news videos where, Fig. 10d is taken from (Ngo and Chan, 2005). Images in Fig. 10e and f are taken from He *et al.* (2005) they are samples of historical archive document images with poor quality due to age and discolored cards. In particular, the characters in Fig. 10a have different colors Fig. 10b is in vertical alignment (c) is double-lined characters and Fig. 10d has complicated background. As global methods, Otsus (1979) and Lin and Yang (2000) methods can not work as well as the adaptive methods are expected. Nevertheless, Otsu's works nicely in Fig. 10e and Lin's results are better than Otsu's in (a). In Fig. 10c, comparing to our result, the binarized lines in "CNN" are not in the uniform width both in Niblack's and Sauvola's. This is because when two adjacent pixels are all text pixels, sliding window centered in one then moving to the next one causes different threshold values. In Fig. 10d, the original image and the binarized results are cited from Ngo and Chan (2005) except ours. It shows that the method in Ngo and Chan (2005) obtains a better result comparing to Niblack (1986) and Wolf *et al.* (2002) but the result of proposed algorithm has almost all the irrelevant background removed. The corresponding intensity information is shown in Fig. 11. The range of background intensity B is (1, 222), leaving the range of foreground intensity F to be (223, 255). Although the width of F is much smaller than B, the threshold T can be determined easily (T = 208, 255) as long as B does not cover the

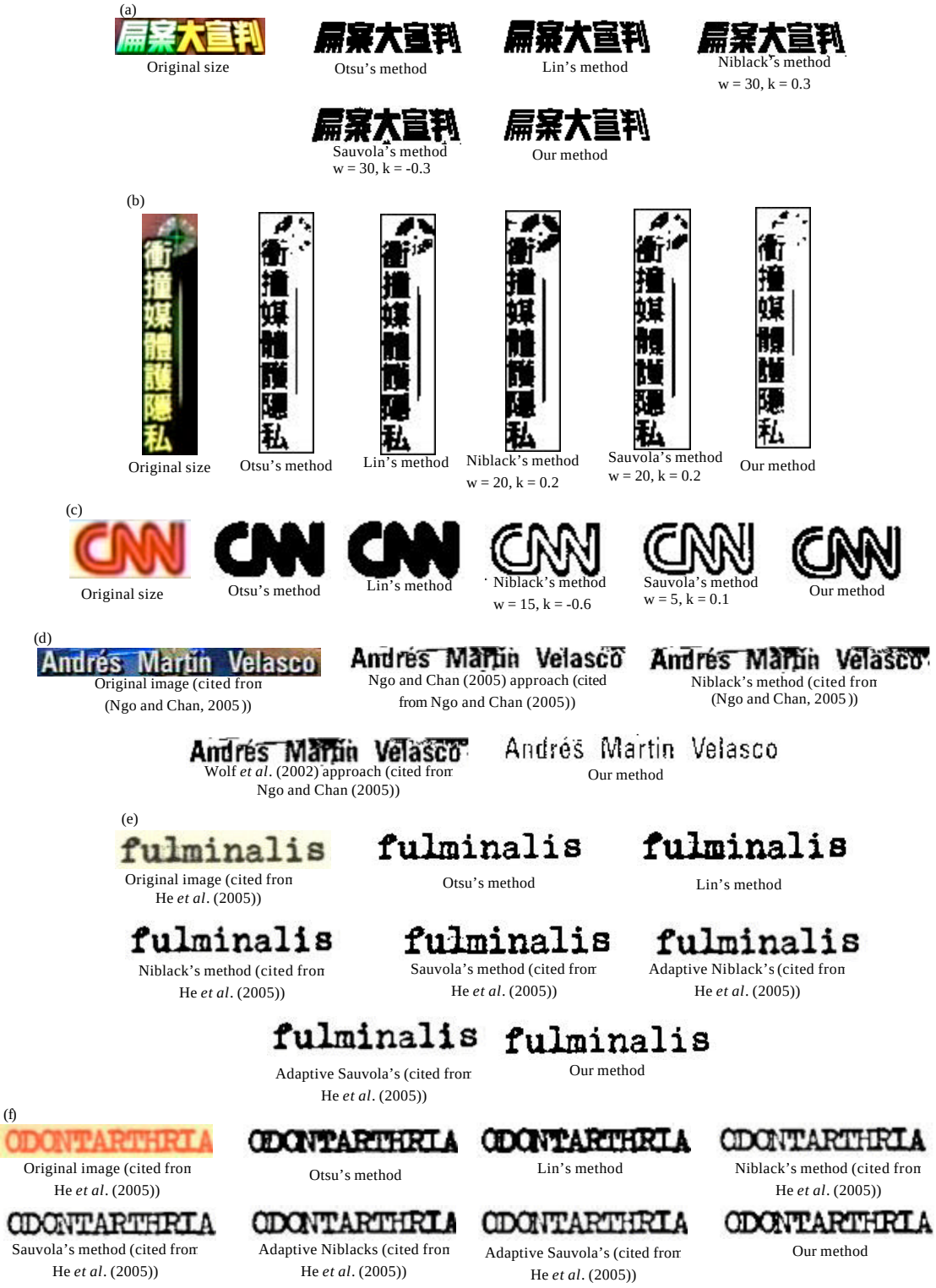

Fig. 10: (a-d) News videos and (e, f) are historical archive document images

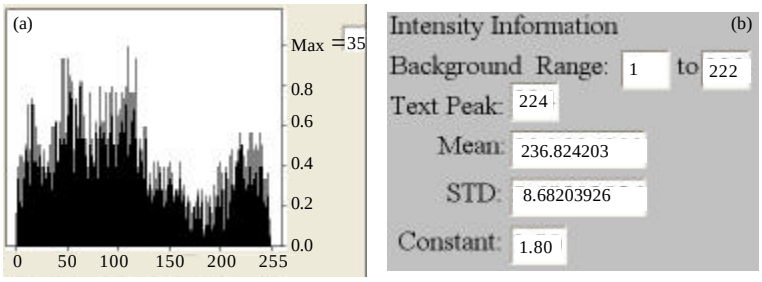

Fig. 11: (a) Histogram and (b) intensity information of the Fig. 10d

entire histogram. In Fig. 10e and f, the documents are in poor quality; Niblack's and Sauvola's results are better than Otsu's and Lin's. The results of the proposed method are clean and precise comparing to the results in He *et al*. (2005) which are improvement of Niblack's and Sauvola's. In these various tests, the proposed algorithm has the best performance which verifies that the proposed method is effective and robust.

CONCLUSION

Global binarization methods can be applied to a windowed or area basis as local binarization and this is exactly the case of the presented work. Similarly, in the application of text binarization on detected boxes, the two global methods, Otsu (1979) and Lin and Yang (2000), can also be viewed in this category. In general, global methods are easier to apply but lack flexibility, opposite to the adaptive methods. This is indeed the case in our experiment with the well-known methods, Niblack and Sauvola. Their results are overall acceptable but the results are at the price of many trials and errors in order to find the most suitable parameters for each case. The proposed method has the merits from both global and adaptive methods. It is parameter-free, computation efficient and robust to various video texts including color, font size and alignment. Combining Canny edge and line-traversing, our method accurately estimates background intensity range and therefore obtaining the correct binarization threshold. Our method can not compute the threshold directly when text intensity is the same as partial background intensity, i.e., the range of background intensity covers the entire intensity range of the image. This situation is not only our problem. It is also a problem of many existing binarization methods. Even if the existing methods can binarize the image, their results are not satisfactory. This problem is solved by integrating the temporal information of video. Using the refined result of the previous work in Yen *et al*. (2010), the image is modified so that background and text have different intensities and thus successfully binarizing the image. Given the modified image to the existing methods, present result confirms the effectiveness. The binarization method also effectively applies to general text in box regions. The algorithm is tested on various degraded document images as well as detected video text boxes in complicated backgrounds. The experimental results outperformed well-known existing methods.

REFERENCES

Gatos, B., I. Pratikakis and S.J. Perantonis, 2006. Adaptive degraded document image binarization. Pattern Recognition, 39: 317-327.

He, J., Q.D.M. Do, A.C. Downton and J.H. Kim, 2005. A comparison of binarization methods for historical archive documents. Document Anal. Recognition, 1: 538-542.

Jung, K., K.I. Kim and A.K. Jain, 2004. Text information extraction in images and video: A survey. Pattern Recognition, 37: 977-997.

Lin, H.J. and F.W. Yang, 2000. An intuitive threshold selection based on mountain clustering. JCIS'2000, New Jersey, USA. http://tkuir.lib.tku.edu.tw: 8080/dspace/handle/987654321/37341.

Lyu, M.R., J. Song and M. Cai, 2005. A comprehensive method for multilingual video text detection, localization and extraction. IEEE Trans. Circuits Syst. Video Technol., 15: 243-255.

Ngo, C.W. and C.K. Chan, 2005. Video text detection and segmentation for optical character recognition. Multimedia Syst., 10: 261-272.

Niblack, W., 1986. An Introduction to Digital Image Processing. Prentice-Hall International, Englewood Cliffs, New Jersey, pp: 115-116.

Ntogas, N. and D. Ventzas, 2008. A binarization algorithm for historical manuscripts. Proceedings of the 12th WSEAS International Conference on Communications, July 23-25, Heraklion, Greece, pp: 41-51.

Otsu, N., 1979. A threshold selection method from gray-level histogram. IEEE Trans. Syst. Man Cybern., 9: 62-66.

Sauvola, J., T. Seppanen, S. Haapakoski and M. Pietikainen, 1997. Adaptive document binarization. Int. Conf. Document Anal. Recognition, 1: 147-152.

Sezgin, M. and B. Sankur, 2004. Survey over image thresholding techniques and quantitative performance evaluation. J. Elect. Imaging, 13: 146-165.

Trier, O.D. and A.K. Jain, 1995. Goal-directed evaluation of binarization methods. IEEE Trans. Pattern Anal. Machine Intell., 17: 1191-1201.

Wang, R., W. Jin and L. Wu, 2004. A novel video caption detection approach using multi-frame integration. Pattern Recognition, 1: 449-452.

Wolf, C., J.M. Jolion and F. Chassaing, 2002. Text localization, enhancement and binarization in multimedia documents. Int. Conf. Pattern Recognition, 2: 1037-1040.

Yager, R. and D. Filev, 1994. Generation of fuzzy rules by mountain clustering. J. Intell. Fuzzy Syst., 2: 209-219.

Yen, S.H., H.W. Chang, C.J. Wang and C.W. Wang, 2010. Robust news video text detection based on edges and line-deletion. WSEAS Trans. Signal Process., 6: 186-194.