2011 International Conference on Document Analysis and Recognition

Binarization of Textual Content in Video Frames

Konstantinos Ntirogiannis[1,2], Basilis Gatos[2], Ioannis Pratikakis[2,3]

[1]National & Kapodistrian University of Athens
Department of Informatics & Telecommunications
GR-15784 Panepistimioupoli, Ilissia, Athens, Greece
kntir@di.uoa.gr

[2]Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-15310 Agia Paraskevi, Athens, Greece
{kntir, bgat, ipratika}@iit.demokritos.gr

[3]Democritus University of Thrace
Department of Electrical and Computer Engineering
GR-67100 Xanthi, Greece
ipratika@ee.duth.gr

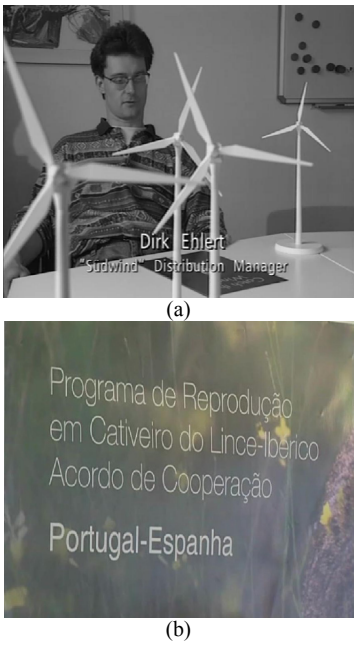
***Abstract*—In this paper we present a binarization technique for textual content in video frames which can be applied in the resulting image of the text detection step aiming in an improved OCR performance. The proposed technique is based on the detection of the text baselines in order to define the main body of the text. The main body of the text is used to detect the stroke width of the characters which will address the two consecutive locally adaptive binarization steps that follow. At the first step, we use different valuation in parameters for the inside and outside area of the main body of the text. To include the thinned or broken binarized parts that may exist outside the main text body, convex hull analysis is performed so that the entire text body is considered. At the second step, binarization is performed with different valuation in parameters for the inside and outside area of the entire text body. The effectiveness of the proposed technique is demonstrated by both qualitative and OCR-based evaluation.**

Keywords - Binarization; video binarization; video OCR

I. INTRODUCTION

Textual information is not only appearing in documents, but it could also be found in video frames. The text is placed over the video scene (synthetic text) in order to offer additional information to the viewers (Fig. 1a). Moreover, textual content may exist as part of the video scene and it is known as scene text (Fig. 1b).

There exist several techniques that perform binarization on the textual content in video frames aiming in an improved OCR performance. Those techniques perform initially a text detection step to detect the text boxes of the video frames using generally edge-based approaches. Focusing on the binarization step for the processing of video frames, many techniques [1-4] incorporate modifications of well-known binarization techniques. In [1], the binarization is performed using the Logical Level Technique of [5] which was developed specifically for document images. Additional steps of [1], that include the morphological operations of opening and dilation, were used to smooth the binarization result of [5]. In [2], the authors used modified 2D Gabor filters to enhance the original grey scale image. Many heuristics were used for the enhancement and the final result was achieved using Otsu [6] on the enhanced grey scale image. It is worth mentioning that in [3], the authors performed the text detection step using an Otsu-based binarization result accompanied by morphological operations. The final binarization was performed with the Sauvola et. al. method [7], using a modified formula for the threshold.



(a)

(b)

Figure 1. (a) Example video frame with synthetic textual content, (b) example video frame with scene text.

Other related works concern the techniques of [8-11], which are mainly based on training. In [9], several binary images representing continuous gray levels of the original image were produced (multi-level thresholding) which were then fed into an SVM classifier. The SVM classifier selected the images that contained text and properly combined them if they were attributed to the same original image. Another work that also required training was described in [10], in which a convolutional neural network was used. The main steps of [10] were convolution, sub-sampling, up-sampling and inverse convolution of the RGB information of the text box. In the most recent related work [11], the Canny edge detector [12] was used

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.141

to specify the text boundaries on the image. Then, a flood fill algorithm was used to fill the edge contour and form the characters. However, the Canny edges can be very confusing since they also depict non-text objects. Especially, in videos with high background complexity the edges of text may connect with background edges and hence deform the actual contour of the characters.

In this paper, we assume that the text detection step has already been performed and we focus on the binarization step of the detected text boxes. We introduce a binarization technique that aims in improving the text/background separation. The main idea is to specify the main body of the text in order to extract valuable information concerning the textual content. The main body of the text is defined as the area which is limited by the upper and lower baseline. Then, within the main body of the text we detect the stroke width (SW) of the characters which is used in consecutive adaptive binarization steps that follow. At a next step, we perform adaptive binarization with different valuation in parameters for the inside and outside area of the main body of the text. Hence, we remove most of the non-text information but in certain cases it results in the thinning and breaking of the textual parts that are outside the main text body. Afterwards, we define the entire text body as the region inside the convex hulls of continuous connected components and we perform adaptive binarization with different valuation in parameters for the inside and outside area of the entire text body. Fig. 2 shows the overall stages of the proposed technique.

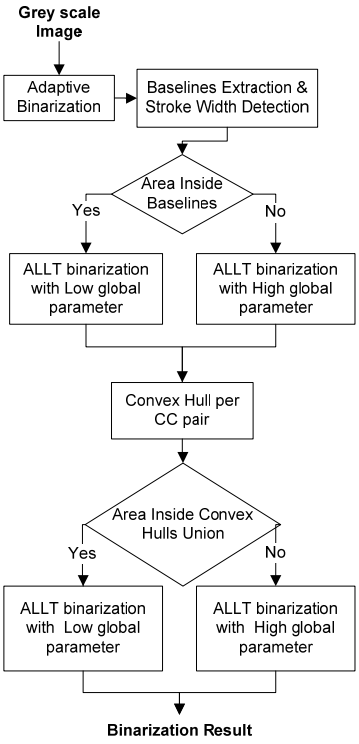


Figure 2. The flowchart of the proposed binarization technique.

II. BINARIZATION GUIDED BY THE TEXT BASELINES

Our initial aim is to detect the upper and lower baselines of the text image. The area which is limited by the lower and upper baseline defines the main body of the text and hence, we can perform binarization with different valuation in parameters for the inside and outside area of the main body of the text. The final result is achieved by extending properly the text region using the convex hulls of the connected components to perform once again the same binarization scheme.

A. Baselines and Stroke Width Detection

For the baselines detection we use the binarization result of an adaptive binarization technique. We use the Modified Adaptive Logical Level Technique [13] with global parameter "a" equal to 0.5 to erase as much background noise as possible. In [13] we had reported a thorough study with respect to the performance of the corresponding technique in relation to the global parameter 'a' and the best performance was achieved with "a" equal to 0.4. Thus, with the chosen value of 0.5 we achieve better results in terms of precision without a significant loss in textual information.

The baselines are detected following the approach in [14], according to which, a linear regression is applied on the set of points that are the lowest/highest foreground pixels in each image column for the lower/upper baseline, respectively (Fig. 3b). Then, the main body of the text is defined as the region between the upper and the lower baseline (Fig. 3c).

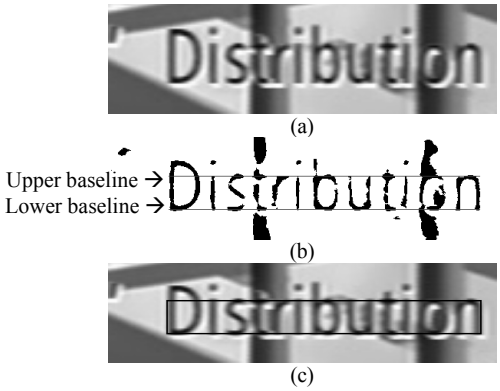


Figure 3. (a) The Original greyscale image, (b) the binarization result of the Modified ALLT [13] along with the baselines in grey, (c) the original greyscale image along with the main body of the text indicated by a bounding box in black.

We calculate the overall stroke width (SW) of the image using the region of the main body of the text (Fig. 3c). Within this region we calculate the average stroke width of the image using only the foreground pixels of the applied binarization [13] along with the stroke width map of [13] (Fig. 4).

(a)

(b)

(c)

Figure 4. (a) The stroke width map of [13]. The contour points of the binarization result of Modified ALLT [13] that is within the baselines area are shown in black, (b) the binarization result of [13] that is within the main body of the text as specified by the baselines, (c) the stroke width map that will be used to calculate the stroke width of the image.

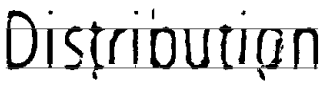


Figure 5. The binarization result of ALLT [15] with different valuation in parameters for the inside and outside area of the main body of the text, respectively. The baselines are shown in grey.

B. Binarization using the Baselines Information

In the sequel, we perform ALLT binarization [15] with the overall stroke width SW that we calculated previously (Section II.A). We exploit the information of the baselines and perform binarization with global parameter 'a' equal to 0.25 inside the baselines area and with global parameter 'a' equal to 0.50 for the remaining image area (Fig. 5). In this way, although the binarization emerges most of the textual information, the portions of text that are cited outside the baselines area may appear as broken or thinned parts. This result motivated the next step of our approach.

C. Definition of the Text Region using Convex Hulls

To overcome the problems introduced by the previous binarization step, we shall encapsulate all the connected components inside a minimum bounding area. In this respect, we calculate the convex hull for each pair of neighboring connected components by a left to right ordering (Fig. 6a). The final Region of Interest (ROI) considered for binarization is the union of all the produced convex hull areas (Fig. 6b).

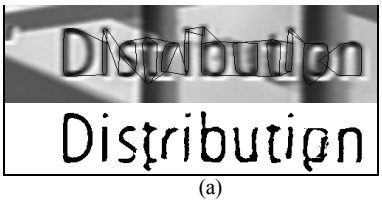
D. Binarization using the Convex Hulls Information

As a final step, we perform the same binarization scheme as in II.B but for the entire body of the text as defined by convex hull analysis (Fig. 7a). The aim is to define more accurately the text body by including all parts of the characters that may be broken or thinned. In the case of using a standard bounding box to define the ROI, much more background and hence more candidate noise is considered (Fig. 7b).

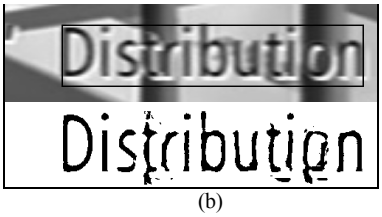
(a)

(b)

Figure 6. (a) The connected components of the previous binarization step along with the convex hulls in grey, (b) the entire body of the text as specified by the convex hulls.



(a)



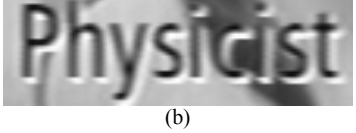
(b)

Figure 7. (a) Binarization result by taking into account the ROI defined by the convex hull analysis, (b) binarization result by taking into account the ROI defined by a standard bounding box.

III. EVALUATION RESULTS

The experiments were conducted using bounding boxes from four video sequences that contained mainly synthetic textual content (Fig. 1 and Fig. 8). The number of bounding boxes used is 44 containing 415 characters that correspond to 62 words.

(a)



(b)

Figure 8. (a) The original video frame, (b) bounding box of the original frame including text with inverted grey values.

The evaluation is performed both at character and word level using the FineReader 8.1 OCR engine [16]. The OCR accuracy [17] concerning the character level is measured according to the following equation:

$$Char\_OCR\_Accuracy = \frac{\#Correct\_Chars}{\#GT\_Chars} \quad (1)$$

where *#Correct_Chars = #GT_Chars - (#Insertions + #Substitutions + #Deletions)*. The OCR accuracy at word level is measured as the ratio between the correctly recognized words and the number of the ground truth (GT) words.

$$Word\_OCR\_Accuracy = \frac{\#Correct\_Words}{\#GT\_Words} \quad (2)$$

For comparison, we tested the OCR performance using several binarization techniques:
(i) ALLT (AL) [15], (ii) Bernsen's method (BER) [18], (iii) Gatos' method (GPP) [19], (iv) Kim's method (KIM) [20], (v) the modified ALLT (MA) [13], (vi) Niblack's method (NIB) [21], (vii) Otsu's method (OTS) [6], (viii) Sauvola's method (SAU) [7].

We also tested the OCR performance of the FineReader [17] using the original grey scale images as input. Moreover, the ALLT [15] technique was tuned for the video images used. The ALLT [15] technique divides the image into blocks and measures the run-length of the blocks that are bimodal to detect the stroke width, taking into consideration the fact that several text lines exist. However, in our case there is only one text line with one or only a few words. Hence, we perform the stroke width detection methodology of the ALLT [15] using all the blocks, i.e. the entire bounding box.

In Tables I-II, the evaluation results of repsesentative cases are shown along with the corresponding OCR output. In Fig. 8a, the entire video frame that concerns the example case of Table II is shown. Lastly, in Table III the overall evaluation results concerning all the text boxes are presented. From Tables I-II we can observe that the proposed technique achieves better binarization and aids the OCR in achieving higher performance.

Additional experimets were performed using word images from the competition in [22]. In this competition the images were captured by a camera and not from video frames. In many cases the text baselines were correctly detected despite the text skew or perspective (Fig. 9). However, in some cases, the text baselines could not be sucessfully detected but the performance of the proposed technique remained satisfactory (Fig. 10). In Fig. 10, the binarization results of the top three techniques according to Table III are presented, as well.

The proposed technique can be applied to both synthetic and scene text but it is more robust for video frames with synthetic textual content. The synthetic text is usually horizontally aligned and the characters have nearly the same intensity. Therefore, the baselines detection is succesful in most cases.

TABLE I. EXAMPLE CASE NO.1

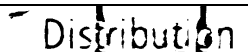
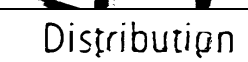
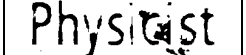
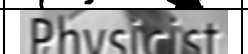
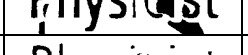
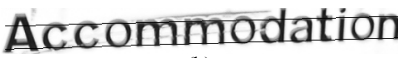
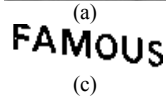
Technique	Binarization	OCR output
AL [15]		Distributes n
BER [18]		TJisffibuvigT
FR [16]		DisffibutiBn
GPP [19]		Distribution
KIM [20]		DisFributjfl n
MA [13]		Distribution
NIB [21]		Dis|ibutjBn
OTS [6]		^sfribuφHi
SAU [7]		DisSribuvji n
Proposed		Distribution

TABLE II. EXAMPLE CASE NO.2

Technique	Binarization	OCR Output
AL [15]		Physitast J «v*
BER [18]		'sltlst
FR [16]		PhysHftt
GPP [19]		Rhys mist
KIM [20]		ist
MA [13]		Physfŏst
NIB [21]		Hhysragt
OTS [6]		Physnatt
SAU [7]		Physftwst
Proposed		Physicist


(a)


(b)


(c)

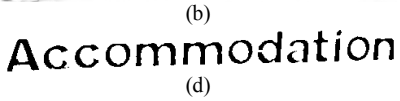

(d)

Figure 9. (a)-(b) Original images from the dataset of [22] along with the baselines, (c)-(d) the binarization of (a)-(b) using the proposed method.

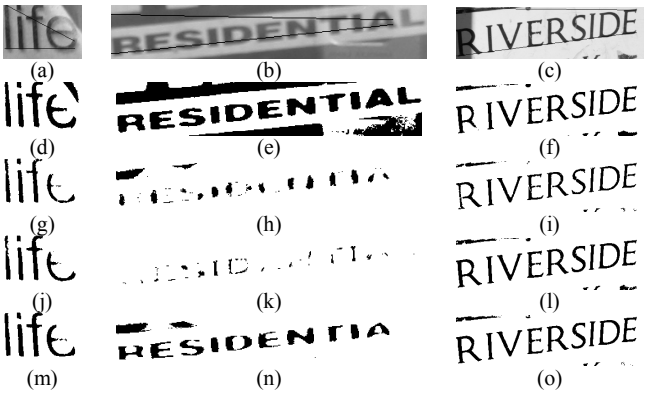


Figure 10. (a)-(c) Original images from the dataset of [22] along with the baselines, (d)-(f) KIM [20] binarization, (g)-(i) MA [13] binarization, (j)-(l) AL [15] binarization, (m)-(o) the binarization results using the proposed technique.

TABLE III. CHARACTER AND WORD ACCURACY USING TEXT BOXES FROM VIDEO FRAMES

Technique	Char Accuracy	Word Accuracy	Average
NIB [21]	60.96	56.45	58.71
BER [18]	66.51	59.68	63.10
OTS [6]	71.08	64.52	67.80
GPP [19]	82.89	70.97	76.93
SAU [7]	83.86	82.26	83.06
FR [16]	88.92	80.65	84.79
KIM [20]	92.29	83.87	88.08
MA [13]	91.81	87.10	89.46
AL [15]	92.05	87.10	89.58
Proposed	93.73	88.71	91.22

IV. CONCLUSIONS

In this work we presented a binarization technique for video frames capable of removing most of the background and interfering noise. We used the baselines of the text to perform adaptive binarization and at a next step we enhanced the binarization output using the convex hulls information. The OCR experimental results along with qualitative evaluation of the results prove the effectiveness of our technique, especially in cases with synthetic textual content.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

REFERENCES

[1] S. Kwak, K. Chung and Y. Choi, "Video Caption Image Enhancement for an Efficient Character Recognition", Proc. 15th Inter. Conf. on Pattern Recgnition (ICPR '00), Sept. 2000, Barcelona, Spain, vol. 2, pp.606–609.

[2] D. Chen, K. Shearer and H. Bourlard, "Text Enhancement with Assymetric Filter for Video OCR", Proc. 11th Inter. Conf. on Image Anal. and Processing (ICIAP '01), Sept. 2001, Palermo, Italy, pp.192-197.

[3] C. Wolf, J-M. Jolion and F. Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Documents", Proc. 16th Inter. Conf. on Pattern Recognition (ICPR '02), Dec. 2002, Quebec, Canada, vol. 2, pp. 1037-1040.

[4] M. Merler and J.R. Kender, "Semantic Keyword Extraction via Adaptive Text Binarization of Unstructured Unsourced Video", Proc. 16th IEEE Inter. Conf. on Image Proc. (ICIP '09), Nov. 2009, Cairo, Egypt, pp. 261-264.

[5] M. Kamel and A. Zhao, "Extraction of Binary Character/Graphics Images from Grayscale Document Images", CVGIP: Comput. Vision Graph. Imag. Process., vol. 55, no. 3, pp. 203-217, May 1993.

[6] N. Otsu, "A Thresholding Selection Method from Gray-level Histogram", IEEE Trans. on Systems, Man and Cybernetics, vol. 9, Mar. 1979, pp. 62-66.

[7] J. Sauvola and M. Pietikainen, "Adaptive Document Image Binarization", Pattern Recognition, vol. 33, no. 2, 2000, pp. 225-236.

[8] C-H. Ngo and C-K. Chan, "Video text detection and segmentation for optical character recognition", Multimedia Systems, vol. 10, no. 3, 2005, pp. 261-272.

[9] J. Li, Y. Tian, T. Huang and W. Gao, "Multi-Pplarity Text Segmentation Using Graph Theory", Proc. 15th IEEE Inter. Conf. on Image Proc. (ICIP '08), Oct. 2008, San Diego CA, USA, pp. 3008-3011.

[10] Z. Saidane and C. Garcia, "Robust Binarization for Video Text Recognition", Proc. 9th Inter. Conf. on Document Anal. And Recognition (ICDAR '09), Sept. 2007, Curitiba, Brazil, pp. 874-879.

[11] Z. Zhou, L. Li and C. L. Tan, "Edge based Binarization for Video Text Images", Proc. 20th Inter. Conf. On Pattern Recognition (ICPR '10), Aug. 2010, Istanbul, Turkey, pp. 133-136.

[12] J. Canny, "A Computational Approach to Edge Detection", IEEE Trans. on Pattern Anal. and Machine Intelligence, vol. 8, no. 6, Nov. 1986, pp. 679-698.

[13] K. Ntirogiannis, B. Gatos and I. Pratikakis, "A Modified Adaptive Logical Level Binarization Technique for Historical Document Images", Proc. 10th Inter. Conf. on Document Anal. and Recognition (ICDAR '09), Jul. 2009, Barcelona, Spain, pp. 1171-1175.

[14] U.V. Marti & H. Bunke, "Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system", Int. Journal of Pattern Recognition and Artifical Intelligence, 15(1), 2001, pp. 65–90.

[15] Y. Yang and H. Yan, "An Adaptive Logical Method for Binarization of Degraded Document Images", Pattern Recognition, vol. 33, pp. 787-807, 2000.

[16] ABBYY (www.finereader.com)

[17] S. V. Rice, "Measuring the Accuracy of Page-Reading Systems". PhD thesis, University of Nevada, Las Vegas, 1996.

[18] J. Bernsen, "Dynamic Thresholding of Grey-level Images", Proc. 8th Inter. Conf. on Pattern Recognition, Paris, France, 1986, pp. 1251-1255.

[19] B. Gatos, I. Pratikakis and S. J. Perantonis, "Adaptive Degraded Document Image Binarization", Pattern Recognition, vol. 39, no. 3, Mar. 2006, pp. 317-327.

[20] I.K. Kim, D.W. Jung and R.H. Park, "Document Image Binarization Based on Topographic Analysis Using a Water Flow Model", Pattern Recognition, vol. 35, 2002, pp. 265–277.

[21] W. Niblack, An Introduction to Digital Image Processing, Englewood Cliffs, NJ PrenticeHall, 1986, p. 115.

[22] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young, "ICDAR 2003 Robust Reading Competitions", 7th Inter. Conf. on Document Anal. and Recognition, Aug. 2003, Edinburgh, Scotland, vol.2,pp.682-687.