This paper will be published on the 13th Int. Workshop on Image Analysis
for Multimedia Interactive Services 23-25 May 2012, Dublin Ireland

A SKELETON BASED BINARIZATION APPROACH FOR VIDEO TEXT RECOGNITION

Haojin Yang, Bernhard Quehl, Harald Sack

Hasso Plattner Institute (HPI), University of Potsdam
P.O. Box 900460, D-14440 Potsdam
email: {Haojin.Yang, Bernhard.Quehl, Harald.Sack}@hpi.uni-potsdam.de

ABSTRACT

Text in video data comes in different resolutions and with heterogeneous background resulting in difficult contrast ratios that most times prohibit valid OCR (*Optical Character Recognition*) results. Therefore, the text has to be separated from its background before applying standard OCR process. This pre-processing task can be achieved by a suitable image binarization procedure. In this paper, we propose a novel binarization method for video text images with complex background. The proposed method is based on a seed-region growing strategy. First, the text gradient direction is approximated by analyzing the content distribution of image skeleton maps. Then, the text seed-pixels are selected by calculating the average grayscale value of skeleton pixels. And finally, an automated seed region growing algorithm is applied to obtain the text pixels. The accuracy of the proposed approach is shown by evaluation.

1. INTRODUCTION

In the past few years, the amount of multimedia content e.g. video data available on the WWW (*World Wide Web*) is constantly growing. Therefore, the analysis and retrieval of video data has become an essential and challenging task. To open up video content for content based search, textual metadata has to be generated either by manual user annotation (e.g. tagging) or has to be extracted by automated analysis.

Although the text in videos can already be localized accurately by excellent text detection algorithms, correct recognition of the identified text via standard OCR tools often does not work with sufficient quality, because standard OCR technology is focussed on high resolution scans of printed (text) documents which comes with black font on white background in properly organized text regions (paragraphs). However, the texts in video images are usually of low resolution and the image background is of heterogeneous complexity (cf. Fig. 1). Therefore, we have to distinguish the text pixels from their background before applying the OCR process to enable high quality results.

In this paper, we propose a new method for video text binarization. Unlike traditional approaches, we utilize image skeleton and boundary maps to detect the text pixels. First, we analyze the distribution of image skeleton maps in order to estimate the text gradient direction, which is necessary to determine whether there is light text on dark background or vice-versa. Subsequently, we calculate the average grayscale value of skeleton pixels, which are extracted in the correct text gradient direction. Making use of this value the seed selection process can be applied followed by a seed region growing algorithm by which seed regions are recursively extended in all four directions until text boundaries are reached. Operability and accuracy of the proposed algorithm have been evaluated by using two publicly available test data sets.

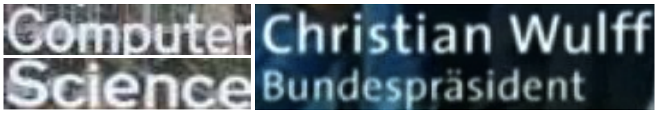

Fig. 1. Texts in video and scene images with a complex background

The rest of the paper is organized as follows. Section 2 presents related work, while in Section 3, the proposed text binarization method is described in detail. Evaluation and experimental results are provided in Section 4. Section 5 concludes the paper.

2. RELATED WORK

Text binarization methods can be categorized into two main groups: thresholding-based algorithms including global thresholding methods [1], which use a single threshold for the entire document and local thresholding methods [2, 3, 4], which assign a threshold for each small local region of the processed document. The second class sums up region-based, clustering-based [5] and edge-based [6] methods.

Chen et al. proposed a multi-hypotheses based framework for video image binarization [7] that applies several existing binarization methods and a standard OCR engine to create the hypotheses for text candidates. The scoring process is performed by using a pre-trained language model. However, this combination of document text binarization methods is not always robust enough when the image has complex background

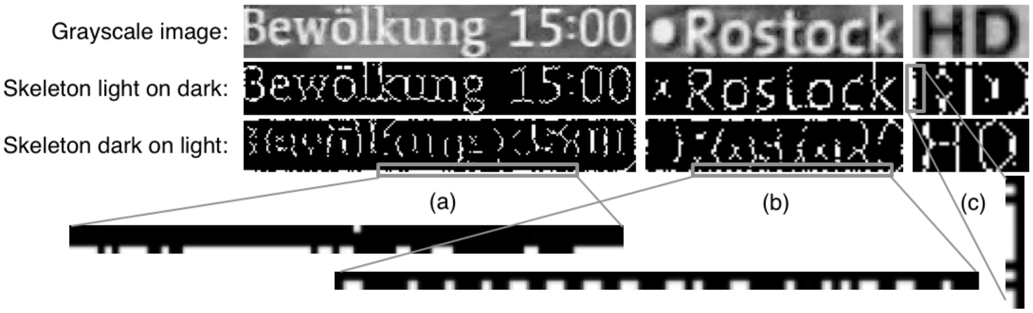

Fig. 2. Text gradient analysis: the first skeleton map is created according to *light on dark* (light text on dark background) rule, whereas the second one is created according to *dark on light* rule.

or low contrast ratio.

Zhou et al. proposed an edge-based binarization method for video text images [6]. The grayscale histogram is used to determine the text gradient direction. The seed selection algorithm is based on the local maximum or minimum of edge values. Although parts of this approach are similar to ours, their system has several shortcomings: First, in order to determine the text gradient direction, their system processes the image histogram with a Gaussian filter recursively until only two peaks are obtained. This process is more computationally complex in contrast to our method. Furthermore, they assume that the background is always the peak with the higher value in the output histogram. However, in practice, we have realized that this assumption is not valid when the detected text bounding boxes from the detection stage are close to the texts. Moreover, our skeleton-based seed selection method has been able to achieve more accurate results, because along the correct gradient direction, the skeleton pixels will be created in the middle of the character shape. In their algorithm, the pixels surrounding the character boundaries have also been considered for the threshold determination, which might have affected the seed selection accuracy.

3. PROPOSED APPROACH

In a standard framework for video OCR, the text binarization process usually follows after text detection and text localization. A text detector determines, whether a video image contains text lines, for which text localization retruns a tight bounding box. In this work, we use the text detection and localization method from [9] as a base line. The binarization is carried out on the detected bounding box images.

3.1. Text Gradient Analysis

To determine the text gradient direction, which is required by our seed selection method, we have tried various image features and methods, such as e.g., histogram of gradient feature [10], variance of stroke width feature [9] etc. However, the best results in our experiments have been achieved by analyzing the content distribution of skeleton maps. Skeleton serves as a thin version of object shapes that represents the structure of objects [11]. We apply the Otsu method [1] to create the binary image which is subsequently adopted for skeleton map computation.

Fig. 2 shows some text bounding box images and their corresponding skeleton maps. By evaluating skeleton maps, which have been created with the wrong gradient direction (cf. skeleton-*dark on light* in Fig. 2 (a) (b) and skeleton-*light on dark* in Fig. 2 (c)), we can see that there are many white pixels placed on the image boundaries. This is due to the characteristics of the skeleton operation. The morphological skeleton can be considered as a special thinning function. Applying the skeleton operation, only the structure of the original shape is preserved. All redundant pixels will be removed. Therefore, along the correct gradient direction the skeleton pixels will be retained in the center line of the character shapes (cf. skeleton-*dark on light* in Fig. 2 (c) and skeleton-*light on dark* in Fig. 2 (a) (b)). Otherwise, the skeleton pixels all surround the characters and are placed on the image boundaries.

Thus, we simply obtain the text gradient direction by comparing the white pixel count on the image boundaries of two skeleton maps. This method has been able to achieve over 95% accuracy for our test data.

3.2. Text Extraction Using Seed-Region Growing

In the previous step, we have determined the text gradient direction. In this section, we describe our seed selection and seed-region growing algorithm.

In order to distinguish text pixels and their background, we apply the skeleton map which has been created with the correct gradient direction to calculate the according thresholds. Let T_h, T_l and T_{mean} denote the highest, lowest, and mean grayscale value of all skeleton pixels. To guarantee that the selected seed pixels are located inside the character boundaries, we have introduced a variance seed factor σ. The

Table 1. Comparison results on German TV news test set

Method	Correct characters	Correct words	Char accuracy	Word accuracy
Yang 2011 et al. [8]	1564	215	0.51	0.44
Otsu [1]	1761	241	0.58	0.49
Niblack [2] (k=0.3)	1822	235	0.59	0.48
Sauvola et. al [3] (k=-0.01)	1653	208	0.54	0.42
Wolf 2001 (k=-0.2) [4]	1717	212	0.56	0.43
Wolf 2007 (k=-0.2) [4]	1627	185	0.53	0.38
Zhou et al. [6]	1925	255	0.61	0.52
Our method	**1997**	**284**	**0.65**	**0.58**

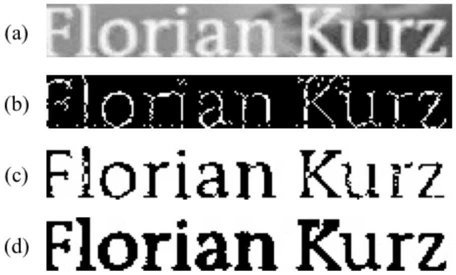

Fig. 3. Seed selection and region growing results: (a) grayscale image, (b) skeleton image, (c) seed image, (d) region growing result.

seed selection procedure works as follows:

if *light text on dark back* **then** $S_p > T_{mean} + \sigma \wedge S_p \leq T_h$
if *dark text on light back* **then** $S_p < T_{mean} - \sigma \wedge S_p \geq T_l$

where S_p denotes the grayscale value of seed pixels. If a pixel satisfies the above conditions, it will be labeled as a seed pixel. For our experiments a seed factor σ=-38 has been determined empirically. Fig. 3 (c) shows the achieved seed selection result.

After seed selection, the process continues with a seed-region growing algorithm. This process determines the pixels to be included in the growing region that forms the text area to be processed by the OCR engine. Seed region growing starts from each seed pixel and extends the seed-region in north, south, east, and west directions. The image edge map is used to terminate the growing process as follows: If the seed-region reaches a edge boundary pixel, this pixel will be labeled as a text pixel, and no further extension will be performed in this direction. To avoid open (not completely closed) text-boundaries, we have applied a similar method as described in [6]: For example, let $P(x, y)$ denote the current seed pixel. To analyze its north neighbor $P(x, y - 1)$, two parallel neighbors $P(x - 1, y - 1)$ and $P(x + 1, y - 1)$ are examined. If one of them is a boundary pixel and the other one is not, then a potential boundary gap has been detected. Then, we label $P(x, y - 1)$ as text pixel and terminate further extension in this direction. For other directions, we repeat the process in the same manner.

Fig. 2 (d) shows the final region growing result that is also the final result of the binarization.

4. EVALUATION AND EXPERIMENTAL RESULTS

The evaluation for our proposed text binarization method is performed on the following two test sets: A set of collected video frames from German TV news program (72 frames, 348 text bounding box images, 490 words, and 3057 characters) that is subsequently referred to as TV news test set [1], and the ICDAR 2003 database ”sample of words” (171 words, 850 characters) [2].

4.1. Video Text Images

We applied the text detection method proposed by Yang et al. [9] in our evaluation as a base line. The achieved pixel based recall and precision of text detection are of 84% and 87%, respectively. The achieved text recognition accuracy in Table. 1 is based on this result. We applied the open-source OCR engine *Tesseract-ocr*[3] for the text recognition.

In order to provide a comparison to other existing text binarization methods, we also applied 7 different reference methods on our TV news test set. In particular, we applied the C++ implementation of the methods *Niblack*, *Sauvola*, *wolf2001*, and *wolf2007* from [4]. In oder to achieve the best result for each method, the optimal parameter k has been manually determined. The results are illustrated in Table. 1. The proposed method outperforms the results of all other reference procedures.

[1] The test data of TV news test set including manual annotation used for this evaluation are available at http://www.yovisto.com/labs/VideoOCR/

[2] http://algoval.essex.ac.uk/icdar/Datasets.html#Robust Word Recognition

[3] http://code.google.com/p/tesseract-ocr/

[4] http://liris.cnrs.fr/christian.wolf/software/binarize/index.html

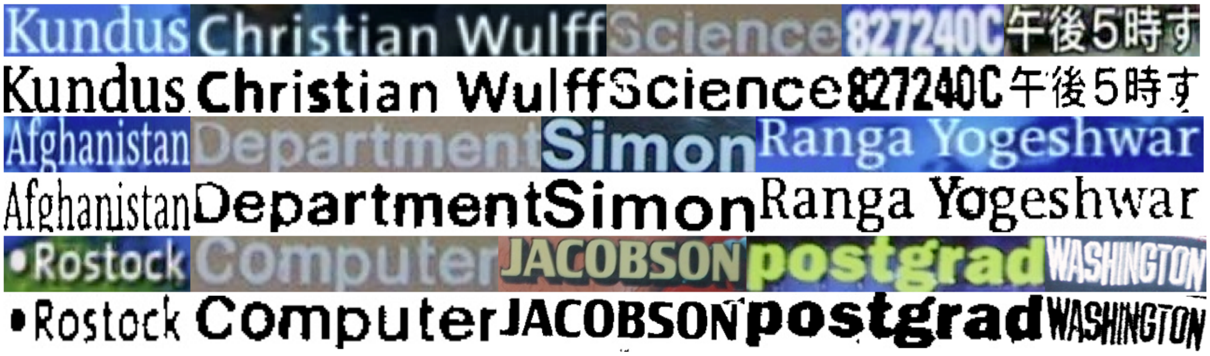

Fig. 4. Some example binarization results of our method from German TV news test set and ICDAR 2003

4.2. Scene Images

In order also to provide a more general evaluation for our method, we additionally performed the evaluation on ICDAR 2003 database ”sample of words”. We have used the evaluation method of ICDAR 2003 robust reading competition, in which the percentage of correctly recognized characters in the ground truth and the achieved experimental results are used to determine recall and precision. The totally recognized characters with our method are 451, of which 365 are correct. The achieved recall and precision are of 43% and 81%, respectively. Since we have not found any other results published for this test set yet, we decided to report our recognition results to establish a new baseline for comparison. Fig. 4 shows some exemplary binarization results of our method from both test set.

5. CONCLUSION

In this paper, we have presented a novel binarization approach for complex video text images. The proposed method consist of three main steps: Text gradient direction analysis, seed pixels selection and seed-region growing. Applying a skeleton map-based analysis we are able to determine the correct text gradient direction. The seed pixels are selected by calculating the average grayscale value of skeleton pixels. After the seed-region growing process, the video text images are converted into a standard OCR engine readable format. Experimental results show that the proposed approach outperforms the other reference methods for recognizing video text images.

6. REFERENCES

[1] N. Otsu, “A threshold selection method from gray level histogram,” *IEEE Transactions on System, Man, Cybernetics*, vol. 19, no. 1, pp. 62–66, 1978.

[2] W. Niblack, *An Introduction to Digital Image Processing*, Englewood Cliffs, New Jersey: Prentice-Hall, 1986.

[3] J. Sauvola and M. Pietikainen, “Adaptive document image binarization,” *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, January 2000.

[4] C. Wolf, J.-M. Jolion, and F. Chassaing, “Text Localization, Enhancement and Binarization in Multimedia Documents,” in *Proc. of the Int. Conf. on Pattern Recognition*, 2002, vol. 2, pp. 1037–1040.

[5] C. M. Thillou and B. Gosselin, “Color text extraction with selective metric-based clustering,” *Computer Vision and Image Understanding*, vol. 107, pp. 1–2, July 2007.

[6] Z. Zhou, L. Li, and C. L. Tan, “Edge based binarization for video text images,” in *Proc. of 20th Int. Conf. on Pattern Recognition*, Singapore, 2010, pp. 133–136.

[7] D. Chen, J. M Odobez, and H. Bourlard, “Text detection and recognition in images and video frames,” *J. of The Pattern Recognition Society*, pp. 595–608, 2004.

[8] H-J. Yang, M. Siebert, P. Lühne, H. Sack, and CH. Meinel, “Lecture video indexing and analysis using video OCR technology,” in *Proc. of the 7th Int. Conf. on Signal Image Technology and Internet Based Systems (SITIS)*, 2011.

[9] H-J. Yang, B. Quehl, and H. Sack, “Text detection in video images using adaptive edge detection and stroke width verification,” in *Proc. of 19th Int. Conf. on Systems, Signals and Image Processing (IWSSIP)*, Vienna, Austria, 2012.

[10] B. Triggs N. Dala, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[11] C.H. Chen, L.F. Pau, and P.S.P. Wang, *Handbook of Pattern Recognition and Computer Vision*, World Scientific, 1993.