H. Yang, B. Quehl, H. Sack:
Text detection in video images using adaptive edge detection and stroke width verification, 19th International Conference on Systems, Signals and Image Processing (IWSSIP 2012), April 11-13, 2012, Vienna (Austria)

TEXT DETECTION IN VIDEO IMAGES USING ADAPTIVE EDGE DETECTION AND STROKE WIDTH VERIFICATION

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 displayed in a video provides important information about the video content. Therefore, it can be utilized as a valuable source for indexing and retrieval in digital video libraries. In this paper, we propose a novel approach for efficient automated text detection in video data: Firstly, we developed an edge-based multi-scale text detector to identify potential text candidates with high recall rate and small computational time expenses. Secondly, candidate text lines are refined by an image entropy based improvement algorithm and a *Stroke Width Transform* (SWT) based verification procedure. Both types of text, overlay and recorded scene text can be localized reliably. The accuracy of the proposed approach is proven by evaluation.

Index Terms— Text detection, video OCR, video indexing, multimedia retrieval

1. INTRODUCTION

In the last decade digital libraries and web video portals have become more and more popular. The amount of video data available on the *World Wide Web* (WWW) is constantly growing. Thus, the challenge of finding video data on the WWW or within digital libraries has become a very important and challenging task. Content-based retrieval within video data requires textual metadata that has to be provided manually by the users or has to be extracted by automated analysis. Techniques from standard *Optical Character Recognition* (OCR), which focus on high resolution scans of printed (text) documents, have to be improved and adapted to be also applicable for video images. In video OCR, first video frames have to be identified that obtain visible textual information, then the text has to be detected and separated from its background and geometrical transformations have to be applied before standard OCR procedures can process the text successfully.

In this paper, we propose a new detection-verification scheme for video text detection. On detection stage, an edge based multi-scale text detector is used to quickly localize candidate text regions with a low rejection rate; For the subsequent verification part, an image entropy based adaptive refinement algorithm can not only reject false positives that expose low edge density, but also further splits all text- and non-text-regions into separate blocks. The then following SWT based verification is adapted to remove the non-text blocks. Operability and accuracy of our text detection algorithm have been evaluated with three different test sets.

The paper is organized as follows. Section 2 presents related work, while in Section 3, the proposed text detection method is described in detail. Evaluation and experimental results are provided in Section 4. Section 5 concludes the paper with an outlook on future work.

2. RELATED WORKS

Most of proposed text detection methods take use of texture features, edges, colors and some text representative features (as e.g., stroke width feature) to discriminate text pixels from background.

DCT *(Discrete Cosine Transform)* coefficents of intensity images have been widely used as texture features for text detection [1, 2, 3]. DCT feature based approach is efficient for JPEG encoded images and MPEG encoded videos. However, in practice this kind of approach can not robustly distinguish text and some text similar textured objects such as bushes or buildings. Therefore, this is unsuitable to handle text detection tasks in complex natural scenes.

Shivakumara et al. [4] proposed a video text detection method based on specialized edge filters for false positive elimination. Unfortunately, their algorithm considered only horizontal text lines.

Recently, some hybrid approaches have been proposed. [5, 6] present two-stage text detection methods, which contain a localization- and a verification-part. For text localization, a fast detector is applied to coarsely detect text, while the verification algorithm refines the results by using *Support Vector Machines* (SVMs). However, this machine learning verification approach can only deliver a binary decision. It might be not robust enough for the bounding boxes from the localization stage that contain both text and non-text pixels.

Epshtein et al. [7] proposed a new image operator SWT

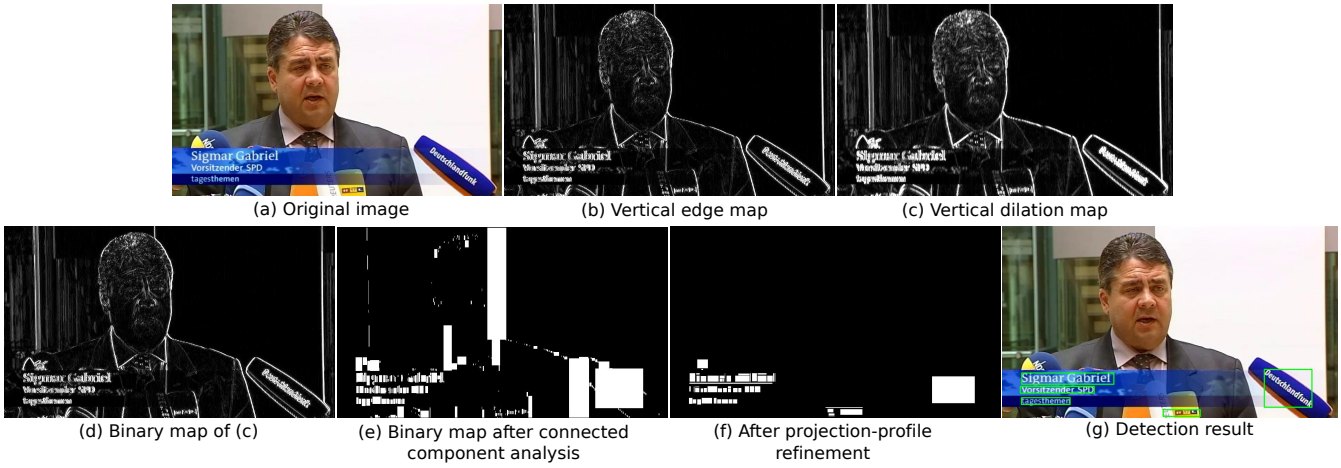

Fig. 1. Workflow of the proposed text detection method. (b) is the vertical edge map of (a). (c) is the vertical dilation map of (b). (d) is the binary map of (c). (e) the result map of subsequent connected component analysis. (f) shows the binary map after the adaptive projection profile refinement. (g) is the final detection result.

for text detection of nature scene images. The operator computes for each pixel the width of the most likely stroke containing the pixel. The output of the operator is a stroke-feature map, which has the same size as the input image, while each pixel represents the corresponding stroke width value of the input image.

3. TEXT DETECTION IN VIDEO IMAGES

Text detection is the first task of video OCR. Our approach determines, whether a single frame of a video file contains text lines, for which a tight bounding box is returned. In order to manage detected text lines efficiently, we have defined a class "text line object" with the following properties: bounding box location (the top-left corner position), bounding box size. After the first round of text detection, the refinement and the verification procedures ensure the validity of the detection results in order to reduce false alarms.

3.1. Text detector

Before performing the text detection process, a gaussian smooth filter is applied to the images that have an entropy value larger than a predefined threshold T_{entr}. For our purpose, T_{entr} =5.25 has proven to be to the best advantage.

We have developed an edge based text detector, subsequently referred to edge text detector. The advantage of our detector is its computational efficiency compared to other machine learning based approaches, because no computationally expensive training period is required. However, for visually different video sequences a parameter adaption has to be performed. The best suited parameter combination of our method were learned from the test runs on the given test data.

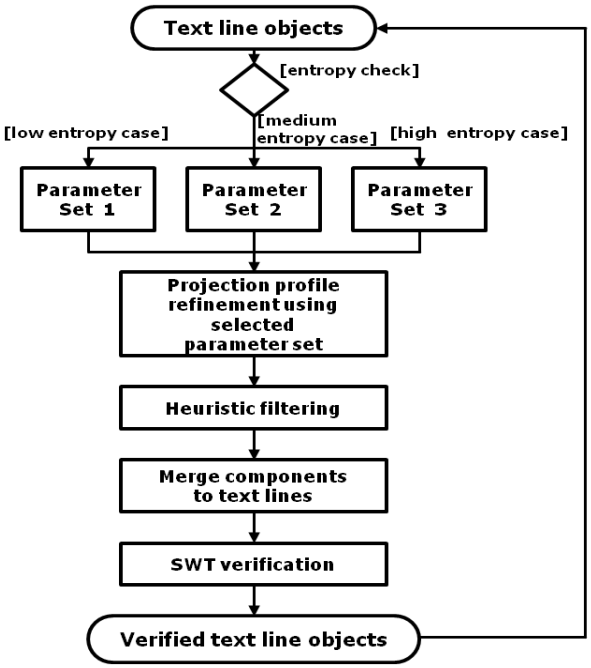

Fig. 2. Workflow of the proposed adaptive text line refinement procedure

The processing workflow for a single frame is depicted in Fig. 1 (a-e). First, a vertical edge map is produced using Sobel filter [8] (cf. Fig. 1 (b)). Then, the morphological dilation operation is adopted to link the vertical character edges together (cf. Fig. 1 (c)). Let $MinW$ denote the detected minimal text line width. A rectangle kernel:$1 \times MinW$ is defined for vertical dilation operator. Subsequently, a binary mask is generated by using Otsu's thresholding method [9]. Ultimately, we create a binary map after *Connected Component*

Table 1. Parameter sets for projection profile analysis. $Edge_{low}$, $Edge_{high}$ denote the thresholds for the *canny* edge filter, whereas min_h, min_v denote the horizontal and vertical minimal edge number, respectively.

	$Edge_{low}$	$Edge_{high}$	min_h	min_v
Parameter set 1	100	200	4	2
Parameter set 2	300	400	5	2
Parameter set 3	500	600	8	2

(CC) analysis (cf. Fig. 1 (e)).

3.2. The text line verification method

We apply a SWT based verification method to candidate text line objects from the previous processing step. Epshtein et al. [7] have proved that the stroke width feature is robust to distinguish text from other complex objects that are visually similar to text such as vegetation. The output of SWT is a feature map, where each pixel contains the potential stroke width value of input image pixels.

In practice, we have found that the computation of SWT performs quite costly for images with a complex content distribution. Moreover, in order to accommodate both bright text on dark background and vice-versa, they apply the algorithm twice for each gradient direction [7]. This also leads to loss of performance. Thus, we only apply SWT on the detected candidate text line images using constraints to verify the text line objects. A text line object is discarded if:

- its stroke width variance exceeds a threshold range $MinVar$ and $MaxVar$,
- its mean stroke width exceeds a threshold range $Stroke_{min}$ and $Stroke_{max}$.

The character components can be generated by merging pixels with similar stroke width value. Furthermore, character components can be merged into character chains, while the mergeable items should have a similar color and a small distance. Thus, a detected text line is discarded, if less than three chains can be detected from it.

In our study, $MinVar$ and $MaxVar$ have been set to 100 and 500, while $Stroke_{min}$ and $Stroke_{max}$ have been set to 1 and 10, respectively. We apply the Euclidean metric to calculate the color distance of *RGB* images. The corresponding threshold value has been set to 40. The threshold value of the character component distance has been set to equal the height of the higher one of the comparison pair.

3.3. Text region refinement

Although the edge text detector has a very high detection rate, many false alarms are created simultaneously. In order to reject falsely detected text line objects, we have developed an adaptive refinement procedure. Fig. 2 shows its workflow. First, the input text line images are classified into three pre-defined categories by using their entropy value. Let E denote the image entropy value, whereas $Entropy_{high}$ and $Entropy_{low}$ denote the corresponding threshold range. The classification is processed as follows:

low entropy case: if $E < Entropy_{low}$
medium entropy case: if $E \leq Entropy_{high} \wedge E \geq Entropy_{low}$
high entropy case: if $E > Entropy_{high}$

The $Entropy_{high}$ and $Entropy_{low}$ have been set to 5.37 and 4.78 in our study, respectively. We have defined 3 parameter sets for the corresponding categories, as illustrated in Table. 1. Where $Edge_{low}$, $Edge_{high}$ denote the threshold values for the *canny* edge filter, while min_h, min_v denote the horizontal and vertical minimal edge number.

Subsequently, the horizontal projection on the edge map of the text line image is processed. A horizontal line is discarded, when its projection value is lower than min_h. In this way, the multi-line bounding boxes are segmented to single lines. Then, we relocate the text line objects to the vertical projection in a similar way. We have adopted the heuristic filter methods to remove objects that are too small or too thin to be readable text. Objects with an invalid aspect ratio will also be removed. The vertical projection might split text lines into single characters. Thus, we have to merge neighbored characters with a distance less than the maximal character height of the according text line.

Lastly, the SWT verification procedure is applied in order to remove the complex false alarms. The refinement process is executed iteratively until no changes occur. Fig. 1(g) shows the result after the adaptive refinement process, which is the final output of the text detection.

Table 2. Comparison results on Microsoft common test set

Method	Recall	Precision	$F_1measure$
Zhao et al. [10]	0.94	0.98	**0.96**
Thillou et al. [11]	0.91	0.94	0.92
Lienhard et al. [12]	0.91	0.94	0.92
Shivakumara et al. [4]	0.92	0.9	0.91
Gllavata et al. [13]	0.9	0.87	0.88
Our method	0.93	0.94	**0.93**

4. EVALUATION AND EXPERIMENTAL RESULTS

The evaluation for our text detection method is performed on three test sets: Microsoft common test set [14], a collected video frame set from German TV news program, and the test

Table 3. Text detection results using pixel based evaluation

	Recall	Precision	$F_1measure$
MS test set	0.92	0.89	0.91
TV news test set	0.86	0.81	0.83
MG test set	0.75	0.81	0.77

set from the Mediaglobe project[1], which are subsequently referred to as MS test set, TV news test set, and MG test set, respectively. Due to copyright restrictions, the videos of the MG test set are not available for public use on the web. However, 104 video test frames including manual annotation used for this evaluation and detailed experimental results are available at [15]. In order to provide a comparison to other existing methods, we applied the evaluation method from [10] that has been proposed for the MS test set. The results are illustrated in Table. 2. However, our proposed method is not able to outperform the results of [10], it surpasses all the other methods for the MS test set. In order also to provide a more general evaluation, we additionally performed a pixel-based evaluation, in which the percentage of overlapping pixels in the ground truth and the achieved experimental results are used to determine recall and precision. The results for all three test sets are illustrated in Table. 3 and show that our approach is also robust enough to work with a wider range of video data than provided in the MS test set.

5. CONCLUSION

In this paper, we have presented a localization-verification scheme for text detection in video images. Our system consists of a fast edge text detector and an adaptive refinement procedure intended to reduce the false alarms. Experimental results show that the proposed method is quite competitive to the other best existing methods. Moreover, our method is also applicable to both Western and Eastern writing systems.

6. REFERENCES

[1] Y. Zhong, H-J. Zhang, and A. Jain, "Automatic caption localization in compressed video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 385–392, 2000.

[2] X. Qian, G. Liu, H. Wang, and R. Su, "Text detection, localization and tracking in compressed video," in *Proc. of Signal Processing: Image Communication*, 2007, pp. 752–768.

[3] H. Yang, M. Siebert, P. Lühner, H. Sack, and C. Meinel, "Automatic lecture video indexing using video ocr technology," in *Proc. of International Symposium on Multimedia (ISM)*, 2011.

[4] P. Shivakumara, T-Q. Phan, and C-L. Tan, "Video text detection based on filters and edge features," in *Proc. of the 2009 Int. Conf. on Multimedia and Expo*. 2009, pp. 1–4, IEEE.

[5] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A two-stage scheme for text detection in video images," *Journal of Image and Vision Computing*, vol. 28, pp. 1413–1426, 2010.

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

[7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. of computer Vision and Pattern Recognition*, 2010, pp. 2963–2970.

[8] I. Sobel, "An isotropic 3x3 image gradient operator," *Machine Version for Three-Dimentional Scenes*, pp. 376–379, 1990.

[9] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SCM-9, no. 1, pp. 62–66, 1979.

[10] M. Zhao, S. Li, and J. Kwok, "Text detection in images using sparse representation with discriminative dictionaries," *Journal of Image and Vision Computing*, vol. 28, pp. 1590–1599, 2010.

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

[12] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 256–268, 2002.

[13] J. Gllavata, R. Ewerth, and B. Freisleben, "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients," in *Proceedings of 17th International Conference on (ICPR'04)*, 2004, vol. 1, pp. 425–428.

[14] X-S. Hua, W-Y. Liu, and H-J. Zhang, "An automatic performance evaluation protocol for video text detection algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 498–507, 2004.

[15] http://www.yovisto.com/labs/VideoOCR/.

[1] Mediaglobe is a SME project of the THESEUS research program, supported by the German Federal Ministry of Economics and Technology on the basis of a decision by the German Bundestag, cf. http://www.projekt-mediaglobe.de/.