Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 11
November 2023

Max-Min Approach for Watermark Text Extraction in Documents

Smt. Maheshwari S. Hiremath[1] and Dr. Basavanna M[2]

[1] Research Scholar, DOS in Computer Science, Davangere University, India

[2] DOS in Computer Science, Davangere University, India.

Abstract:

Online digital content authentication and copyright protection is a crucial issue. A Digital watermarking concept offers a comprehensive authentication and copyright protection for online documents. In contrast to images, audio, and video, text is the most common form of Internet communication and needs to be completely protected. Digital watermark text is a process of adding the watermark text information on online documents to protect from unauthorized access. The extraction of watermark from documents helps to know the quick information about respective document. In this research work the max-min clustering algorithm is employed to increase the semantic gap between the watermark text pixels and background text pixels. Detection process determines the watermark text and actual location of the watermark text in the documents. Thresholding system and post processing heuristic technique is employed to extract the actual watermark text region from the image. The results depicts that proposed model is promising and encouraging.

Keywords: Max-min approach, digital water mark, text detection, sliding window, thresholding.

1. Introduction:

The internet is used to transport a vast amount of information since technology is developing so quickly. For the creators of the material and the owners of the rights, it leads to the illegal interception and copying of verified text documents and resulting in considerable financial losses. The ownership of information and integrity of text documents must be preserved in a method that cannot be seen by everyone. As a result, it becomes necessary to create a well-secured watermarking approach to protect text documents from these crimes. Many scholars have already given text watermarking serious thought, but copyright protection of PDF document has not. A complete copyright protection and authentication solution for digital contents is provided by digital watermarking. A digital watermark is a unique identifier that is permanently inscribed in the data and can be either visible or invisible, with the latter being preferred. Digital text watermarking is the process of including information specific to the document's author or copyright holder in a digital watermark that is embedded into a digital text document. Text watermarking techniques [1, 2] that are effective can prevent copyright breaches. Some of the salient characteristics of plain text that must be handled in any text watermarking method are its binary nature, work patterns, text meaning, grammar structure, writing styles, and language norms.

It can take a long time to manage massive collections of scientific documents. Numerous systems have been created to automatically extract metadata, such as title, authors, and journal, to simplify the management process. However, besides the metadata, it is also helpful to extract the content for in-depth browsing and content searching in digital libraries. Additionally, selected content can be easily shown without scrolling across the Internet and transferred more quickly. However, in addition to these, text mining tasks utilizing machine learning algorithms can also be carried out using the material. When training the machine learning algorithm, having more content is generally preferred. Having a technology that could automatically extract content in large quantities is therefore beneficial.

The text watermarking based on the natural language processing modifies the sentence structure in the original document, embeds the watermark in the new sentences, and alters the substance of the document [3, 4].One of the methods that hold promise for achieving the overall objective of safe document transmission from its source to authorized end users is data hiding. A

multimedia object is "watermarked" by embedding data known as a watermark, tag, or label so that it can be recognized or extracted later to support a claim about the object.

These days, a lot of scientific documents are created in PDF format. However, PDF is designed primarily for viewing and does not retain a document's structural information. It merely comprises of where each character is located on the page. As a result, information extraction from a PDF can be noisy, and the characters that are recovered might not be in the proper order. Additionally, a watermark in a PDF file, which is regarded as an object, can obstruct the sequence of information extraction. A strategy to extract material from PDF with and without watermark in bulk is suggested as a solution to the aforementioned issue. It used optical character recognition (OCR) and direct text extraction from PDF to produce two versions of the digital text [5].

2. Literature Survey:

To the best of our knowledge, there are currently no works that concentrate on extracting content from PDF with a watermark in bulk, despite the fact that there are works connected to extracting content from PDF.

In this area most of the works are done on embedding of watermark text or image on documents, but there are only countable works have been done on extraction and detection of watermark text in images and documents. Fahmy, [6] introduced the quasi blind watermark extraction technique, which requires only a very few information about the original host image that is needed for the complete recovery of both the host image and watermarking logo.

Kim et al. [7] introduced a brand-new learning-based strategy for extracting CNN-based watermark information. They employ a template generating network, which can be found in the template extraction network, to produce a particular type of noise. The extraction stage begins with the extraction network extracting the template, which is then used to evaluate the RST parameters of the geometrically warped image. In the end, the watermark can be decrypted using the estimated watermark matrix.

Yu et al. [8] used the Independent Component Analysis idea to find and extract the watermarking data. The presence and absence of the original images are used to extract the watermark information. In order to recognise and extract the watermark information, the private watermarking system's watermark recovery procedure takes into account both key and original data. This watermark extraction approach isolates the watermark components using a minimum of three linear mixtures, such as the key picture, the original image, and mixed images.

To recognize the watermark text on PDF documents, two different types of text extraction approaches have been proposed by Chai et al. [9]. There are six steps in the suggested method. The user can then choose which portion of the PDF's text to extract after the user's PDF pages have first been converted to photos. In order to convert picture text into digital text, optical character recognition (OCR) is used. At the same time, PDFMinerb is enforced on the PDF in order to extract digital text. The results show that this method is effective for removing text from PDF files that include different types of watermarks.

To validating and confirming the righteousness of watermarked text documents, an Eigen value-based watermark creation approach was developed by TamilSelvan et al. [10]. A secret key, or variable K, controls the watermark in a sensitive technique. The received document is first read, with the image of the document being separated into 8-bit planes. The watermark is then extracted from each bit plane and transformed into a binary form. Connect them all together in the end so you can use a secret key to analyse them.

Mehta et al. [11] proposed a technique to extract the watermark from allegedly malicious document files, which converts the document into pages. A written document is divided into blocks with and without texture using an energy-based method. When the sizes of the original cropped page and the attacked cropped page don't match, the block wise embedding method can result in incorrect block recognition and loss of embedded data. To avoid losing information, the bilinear interpolation attacked cropped page is scaled to the size of the original cropped page. The pages are then divided

into texture and non-texture chunks then divided into dimension blocks. The suggested method extracts the watermark from the texture blocks by using the same watermarking technique that is used for embedding watermark.

Du et al. [12] developed a text digital watermarking algorithm based on a minor alteration in text color.After altering the low 4 bits of the RGB color components of the letters, watermarks were implanted.It substitutes the words in the text with synonyms to implant secret data without altering the sense of the text. With this technique, a document's quality is diminished, and a sizable synonym dictionary is required.

Podilchukand Zeng[13] introduced the IA-DCT technique to extract the partially watermarking decompressed JPEG bit streams. Based on frequency sensitivity and several resolutions, the IA-W system, utilizes the visual models, and image adaptive watermarking techniques for compression purposes. We explored a block-based discrete cosine transform multi-resolution wavelet framework.

Singh et al. [14] discussed the solution to copyright protection, content authentication, and safe ownership of digital photographs is digital watermarking. LSB watermarking, this uses watermarking in both the spatial and frequency domains. Watermarking is the method of embedding a signal, while the other signal serves as the host or cover. Additionally, LSB performed noise cropping on grayscale photos.

Tian et al. [15] used a two-step clustering approach to separate neighboring text instances from the predicted centre and full segmentation maps, treating each text instance as a cluster. Basavaraju et al. [16] developed a level set algorithm to identify the text region and the Gaussian mixture model helps to separate the each character in the image.

Shivakumara et al. [17] applied a Laplacian filter on the input image before using the K-mean classifier to locate text candidates based on the largest difference. The text strings were separated from one another using the skeletons of each connected component. Finally, they used the text string edge's density and length smoothness to eliminate the portions that had been incorrectly detected. This approach performs well at the text line level but poorly at the word level.

Chen et al. [18] created a method that effectively detects the characters in the photos using the directional gradients' better histogram feature. They strengthened the oriented gradient feature's histogram to withstand scale and translation adjustments. The scale and translation robust HOG (STRHOG) acronym was given to this new functionality.

Basavaraju et al. [19] applied hidden Markov random field with EM algorithm to extract the complex text in the image. The text pixels are grouped separately by studying the random variable of a pixel. Shivakumara et al. [20] proposed the gradient vector flow (GVF) to identify the text index pixels in the edge image. Finally, they extracted the final components by grouping and eliminating the non-text components after extracting the edge components that correspond to the appropriate pixels in the edge image.

Basavanna et al. [21] developed a novel run-length based technique for multi-oriented text identification in scene images. To calculate run-lengths for multi-oriented text images, it takes into account one ideal Sobel edge image of the horizontal text image. The boundary expanding approach explores a novel idea based on zero-crossing to distinguish text lines from touching text lines.

Basavanna et al. [22] used sliding window across a word of a text line identified by the technique based on adaptive histogram analysis. The premise for the histogram analysis is that each sliding window's text pixels have uniformly colored intensity values. The technique divides words into segments based on region growth, which examines character and word space.

Basavanna et al. [23] described a novel approach for scene text recognition in complex background images based on the run length algorithm. For the input image's Sobel edge map, run lengths are determined. Due to the consistent spacing between characters and words when text appears in an image, the run length aids in the text detection process.

3. Proposed Methodology:

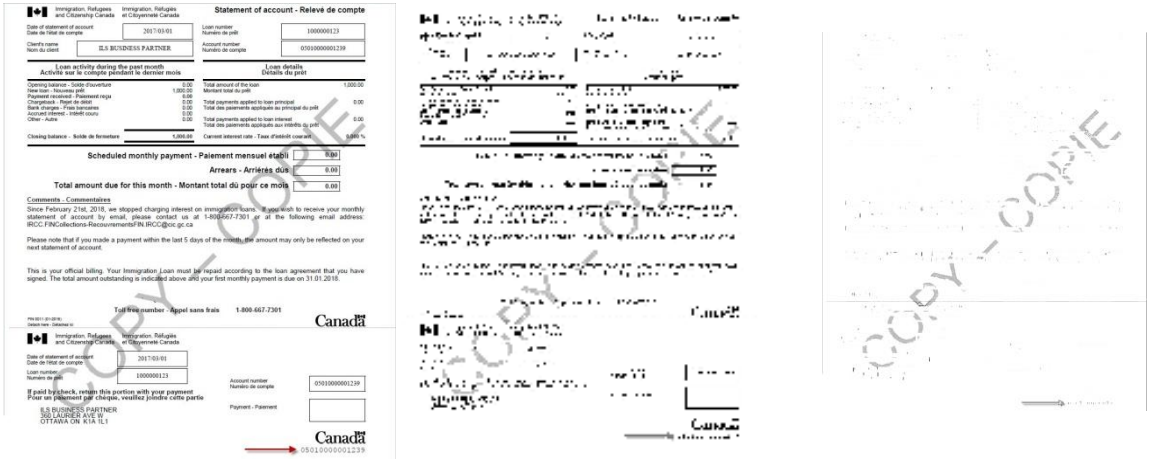
Due to the fact that watermark text pixels have different contrast value than background pixels, it is common to note that the watermark textpixel values can be categorized into three sub bands. In order to extract such an observation, we suggest using Max-Min clustering on three values of each sub band. The resulting image is referred to as an enhanced image since it has varying different values for watermark text pixels and background pixels. As a further measure to isolate the watermark text information, the sliding window Max-Min filtering is again applied along with heuristic rules.

3.1Max-Min Filtering Approach for Watermark Text Enrichment:

In this work, background information is suppressed in the suggested model in order to improve the watermark text information. The rationale behind this grouping is because watermark text pixels differ from background pixels in terms of pixel values. The approach produces three sub-band pictures for the input image in Figure 1(a), which has watermark text with a background. We employ a straightforward Max-Min clustering criterion that chooses the Maximum (Max) and Minimum (Min) value from three sub-bands for each pixel in order to identify the fluctuating watermark pixel value in those bands. To get the value that is closest to the third, the Max and Min values are compared. If the third value is closer to the maximum value, it will form a cluster with the maximum value; otherwise, it will build a cluster with the minimum value. If the third value forms a Max cluster, the procedure selects the Max value within the Max cluster to replace the actual pixel value; otherwise, it selects the Min value within the Min cluster. The Max-Min clustering in this propose method aids in grouping to detect the watermark textpixel value in the input image, resulting in an enhanced image as seen in Figure 1(b). Where one can note that watermark text pixels are enhanced and background pixels are suppressed.

3.2 Grouping of Neighborhood pixels for Watermark Text Detection:

Watermark text pixels in the input image have distinct pixel values, as demonstrated in the previous step. This hint prompts us to use the same grouping criterion to enhance the watermark text while also examining the values of the neighboring text pixels in the enhanced image.As shown in Figure 1(c).Where the watermark text pixels are still brighter than the pixel in the enhanced image shown in Figure 1(b), it is suggested that a sliding window 3x3 operation where we use the aforementioned process to sharpen the watermark text pixels and to suppress the background pixels based on neighbor information further. An experimental investigation is used to estimate the sliding window's size.



(a). Input Image (b). Enhanced Image (c). Sharpen Image

Figure 1: Intermediate results of Max-Min cluster

3.3 Heuristic rules for Watermark Text Separation:

Watermark text pixel values differ in large gap between other text pixels and background pixels of the respective document. With the observation of this evidence, the intensity values of watermark text pixels are retained by applying a threshold system and post processing technique. The processed image is complimented to represent

watermark text pixels as foreground information and remaining other component and non-watermark text pixels as background information. Finally, some of the false positives are eliminated by processing the heuristics rules to obtain the actual watermark text pixels of the given document. Figure 2 shows the final outcome of the proposed methodology.

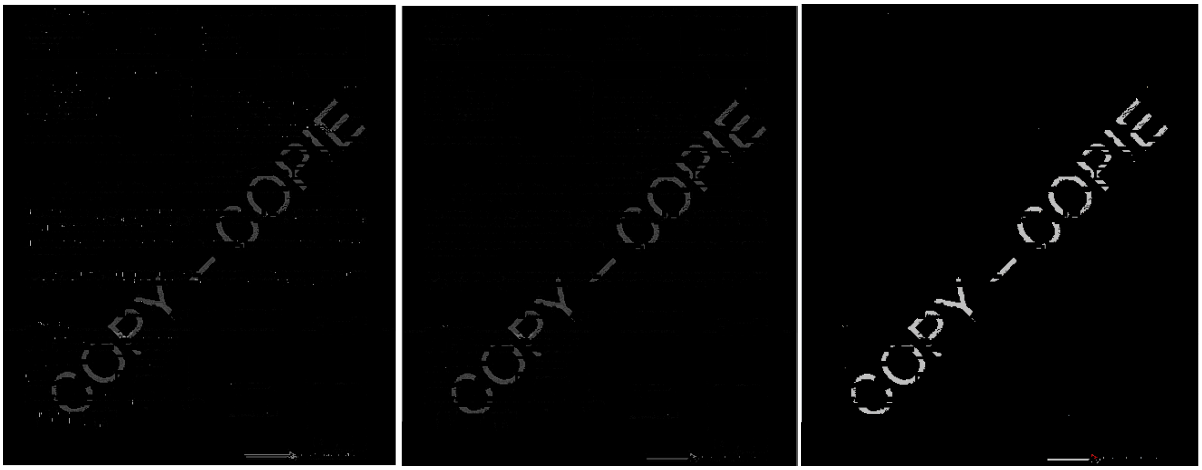


Figure2: Final outcome of the proposed methodology

4. Experimental Analysis:

An experiment has been conducted on a variety of watermarked document images to analyze the performance of a presented watermark text extraction model. The dataset is collected by capturing the watermarked PDF documents and some of the watermark text samples are gathered from the online. The collected dataset contains a number of challenges like variations in color, size, contrast, and orientation. The max-min clustering approach helps to increase the gap between watermark text pixels and background pixels. Finally, heuristic rules are employed to identify the actual watermark text pixels from the given input samples.

The efficiency of suggested method is evaluated by considering the measurements like precision, recall and f-measure. The main factors need to be analyzed to calculate the considered measurements such as Actual Watermark Text Information (AWTI): It consider all watermark text information present in the given input. Truly Detected Watermark Text Information(TDWTI): It takes only the watermark text information identified by the proposed algorithm. Falsely Detected Watermark Text Information (FDWTI): it represents the non-watermark text information is identified as watermark text information.

From the above parameters like AWTI, TDWTI and FDWTI, the precision is calculated as depicted in equation 1, recall is calculated as like equation 2 and f-measure is calculated as represented in equation 3.The max-min cluster effectively separates the watermark text pixels region and background pixels region. The proposed model achieves 84.21% of precision, 91% of recall and 87.47% of f-measure. A sample output of the proposed model is represented in the Figure 3.

$$Precision = \frac{TDWTI + FDWTI}{AWTI} \quad (1)$$

$$Recall = \frac{TDWTI}{AWTI} \quad (2)$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

Before Image

After images perfromed by the proposed method

Before Image

After images perfromed by the proposed method

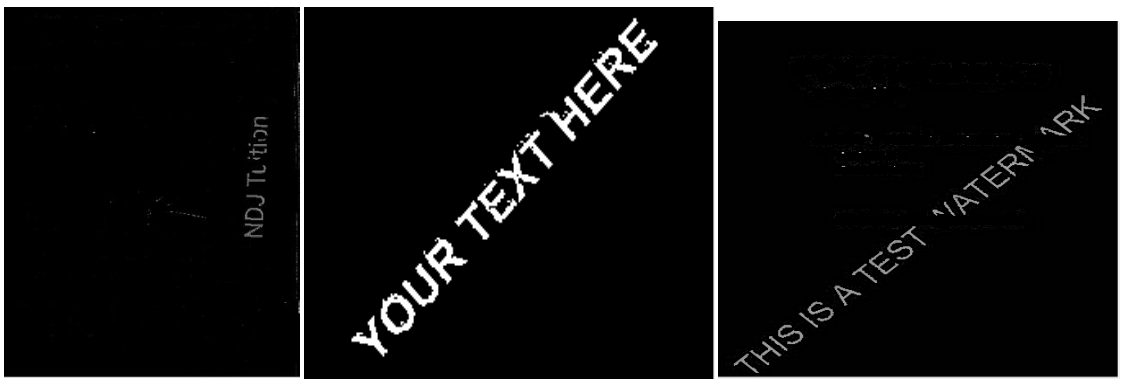


Figure 3: Sample outcomes of the proposed methodology

A new enhancement method developed in [24] for detecting text pixels in videos by performing Laplacian and Sobel operations. The text candidates are obtained by intersecting the Bayesian classifier output with canny edge map. The method of boundary growing is based on the concept of nearest neighbors. The method's resilience has been evaluated on a range of datasets, including own dataset. The text regions are detected using multiple frame enhancement technique [25] by strengthen the contrast between text and non text pixels. A connected component method is implemented to segment actual text regions. The quantitative results reported in Table 1 represents that the proposed method is good for watermark text identification in terms of measures. Watermark text identification process is challenging task due to its own variant features as compared to normal text identification in the images. Most of the time watermark text is in lower intensity level and diagonal orientation with variant text styles. The proposed model extracted the watermark text information from the document efficiently and state of the art technique is depicted in Table1.

Table1: Comparative analysis on existing methods

Methods	Recall	Precision	F-Measure
Basavanna et al. [23]	81	64	71
Shivakumara et al. [24]	87	72	78
Zhou et al.	66	83	73
Proposed Method	91	84.21	87.47

5. Conclusion:

The vast volume of document images is dispersed across the Internet due to the e-commerce and e-government industries' quick expansion. It is vital to safeguard document image copyright and prevent unauthorized transfer and use. In order to address concerns with authentication and information security, watermark extraction is crucial.The Max-Min clustering approach helps to group the watermark textpixel values separately from the background pixel values. And sliding window concept aids to sharpen the watermark text information. Finally, heuristic rules were performed to extract the actual watermark text region. In future work, the statistical measurements will be studied to distinguish the watermark text pixels from the document.

6. Reference:

[1] Jalil Z, Mirza AM (2009) A review of digital watermarking techniques for text documents. In: Proceedings of the International Conference on Information and Multimedia Technology, pp 30–4.

[2] Ranganathan S, Ali AJ, Kathirvel K, Mohan KM (2010) Combined text watermarking. Int J ComputSci Inform Tech 1, 414–6.

[3] Alattar AM, Alattar OM (2004) Watermarking electronic text documents containing justified paragraphs and irregular line spacing. In: Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VI vol 5306, pp 685–94.

[4] Kim Y, Moon K, Oh I (2003) A text watermarking algorithm based on word classification and inter-word space statistics. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), pp 775–9.

[5] Huang D, Yan H (2001) Interword distance changes represented by sine waves for watermarking text images. IEEE Trans CircSyst Video Tech 11, 1237–45.

[6] Fahmy, G. (2009, December). A quasi blind watermark extraction of watermarked natural preserve transform images on a DSP board. In 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 259-264). IEEE.

[7] W.H. Kim, J.U. Hou, S.M. Mun, and H.K. Lee, Convolutional neural network architecture for

recovering watermark synchronization. arXiv preprint arXiv:1805.06199, 2018.

[8] Yu, D., Sattar, F., and Ma, K. K. (2002). Watermark detection and extraction using independent component analysis method. EURASIP Journal on Advances in Signal Processing, 2002(1), 1-13.

[9] Chia, W. C., Teh, P. L., and Gill, C. M. H. D. (2018, July). Text Extraction and Categorization from Watermark Scientific Document in Bulk. In 2018 3rd International Conference on Computational Intelligence and Applications (ICCIA) (pp. 47-51). IEEE.

[10] TamilSelvan, R., Prathap, I., Ramalingam, A., and Raghavan, S. (2009, June). A novel approach to watermark text documents based on Eigen values. In 2009 International Conference on Network and Service Security (pp. 1-5). IEEE.

[11] Mehta, S., Prabhakaran, B., Nallusamy, R., and Newton, D. (2016). mPDF: Framework for Watermarking PDF Files using Image Watermarking Algorithms. arXiv preprint arXiv:1610.02443.

[12] Du M, Zhao Q (2011) Text watermarking algorithm based on human visual redundancy. Adv Inform Sci Service Sci 3, 229–35.

[13] C. I. Podilchuk and W. Zeng, "Image-Adaptive Watermarking UsingVisual Models," vol. 16, no. 4, pp. 525–539, 1998.

[14] A. K. Singh, N. Sharma, M. Dave, and A. Mohan, "A NovelTechnique for Digital Image Watermarking in Spatial Domain," pp.497–501, 2012.

[15] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4234–4243.

[16] Basavaraju, H. T., Aradhya, V. M., Pavithra, M. S., Guru, D. S., &Bhateja, V. (2021). Arbitrary oriented multilingual text detection and segmentation using level set and Gaussian mixture model. Evolutionary Intelligence, 14, 881-894.

[17] Shivakumara P, Phan TQ, Tan CL (2010) A laplacian approach to multi-oriented text detection in video. IEEE Trans Pattern Anal Mach Intell 33(2):412–419.

[18] Chen J, Zhao H, Yang J, Zhang J, Li T, Wang K (2017) An intelligent character recognition method to filter spam images on cloud. Soft Comput 21(3):753–763.

[19] Basavaraju, H. T., Aradhya, V. M., & Guru, D. S. (2019). Text detection through hidden Markov random field and EM-algorithm. In Information Systems Design and Intelligent Applications: Proceedings of Fifth International Conference INDIA 2018 Volume 1 (pp. 19-29). Springer Singapore.

[20] Shivakumara P, Phan TQ, Lu S, Tan CL (2013) Gradient vector flow and grouping-based method for arbitrarily oriented scene text detection in video images. IEEE Trans Circuits Syst Video Technol 23(10): 1729–1739.

[21] Basavanna, M., Shivakumara, P., Srivatsa, S. K., & Kumar, G. H. (2012). Multi-oriented text detection in scene images. International Journal of Pattern Recognition and Artificial Intelligence, 26(07), 1255010.

[22] Basavanna, M., Shivakumara, P., Srivatsa, S. K., & Kumar, G. H. (2016). Adaptive Histogram Analysis for Scene Text Binarization and Recognition. Malaysian Journal of Computer Science, 29(2), 74-85.

[23] Basavanna, M., Shivakumara, P., Srivatsa, S. K., & Kumar, G. H. (2011). A New Run Length based Method for Scene Text Detection. In IICAI (pp. 1730-1736).

[24] P. Shivakumara, R. P. Sreedhar, T. Q. Phan, S. Lu and C. L. Tan, Multi-oriented video scene text detection through Bayesian classi cation and boundary growing, IEEE Trans. CSVT 22 (2012) 12271235.

[25] J. Zhou, L. Xu, B. Xiao and R. Dai, A robust system for text extraction in video, in Proc. ICMV (2007), pp. 119124.