A Combinational approach towards Multioriented Videotext Detection and Recognition

P.Sudir, M.Ravishankar

Abstract—Video text detection, extraction and recognition is an important prerequisite for the high-level semantic information understanding and content-based Video analysis tasks.Inevitable computer vision based loopholes such as cluttered background, inconsistent colors and arbitrary orientations pose interesting challenges in video text content extraction and understanding.A Novel approach towards detection and recognition of Multioriented Videotext is proposed in the present paper. Initially fast and effective Morphological procedures are applied for text region detection followed by Maximally Stable Extremal Regions(MSERs) extraction procedure.MSER detected Text regions are further verified by Covariance Matrix feature extraction based method.Also a novel Watershed and Threshold based videotext binarization method for Video text recognition is proposed.Operability and accuracy of proposed system are evaluated on the standard datasets results which successfully confirm the efficiency of the proposed method.

Index Terms— Covariance Matrix, Morphology, MSER, Scene Text, Video text, Videotext binarization, Watershed Transform.

1 INTRODUCTION

In Recent times Video content analysis and understanding has been greatly facilitated by Text detection and recognition procedures in videos.Since videotext gives a clear description of the instances in videos,the text recognition procedure has become an integral part of video retrieval system.Design of efficient Optical Character Recognition(OCR) systems directly related to Video documents is an important issue.With the arrival of numerous vision related applications on mobile devices including the iPhone and Android platforms,which can translate text into other languages in real time,has amplified the interest in the problems[1], [2].However,the huge diversity of texts in terms of difficult acquisition conditions (low resolution,complex background,non uniform lighting, occlusions, and blurring effects) make the task of text detection and recognition a challenging problem that has raised a growing interest in recent research activities[3], [4].Last three decades have seen numerous efforts in addressing the hitches faced during text area detection[5], [6]text tracking[7], [8] and text recognition[9], [10], [11].

Advances towards text detection approaches can be classified into three categories.The first category are Color and Connected Component (CC) based methods [12], [13]. Color layers are generated by using color reduction techniques.Various clustering algorithm are employed to obtain CCs and connected CCs are classified into text candidates with color similarity and component layout analysis. Karatzas and Antonacopoulos [14] extracted text components with a split-and-merge strategy in the hue-lightness-saturation (HLS) colorspace. Yi et al.[15] proposed a new bigram color uniformity based method to model both text and attachment surface and cluster edge pixels based on colorpairs and spatial positions into boundary layers.Then,stroke segmentation is performed at each boundary layer by color assignment to extract character candidates.

Although Color and CC based methods can locate text quickly they fail in case of text which are embedded in complex background or touching other graphical objects.The second category is Edge based methods[16],[17],[18] which detect text regions by analyzing the edge intensity maps.In recent work Phan et al.[19] proposed a text detection approach which groups horizontally aligned components of ″gradient vector flow″ into text candidates.Gradient vector flow is used to extract both intra-character and intercharacter symmetries.Finally to remove false positives,a learning-based approach is employed which uses Histogram of Oriented Gradients feature.Huang et al[20]. proposed a new edge-based method which uses an edge-ray filter, to detect the scene character.A method is proposed to filter out complex backgrounds by fully utilizing the essential spatial layout of edges instead of the assumption of straight text line.Edges are extracted by a combination of Canny and Edge Preserving Smoothing Filter (EPSF).A new Edge Quasi-Connectivity Analysis (EQCA) is employed to unify complex edges as well as contour of broken character.However,these kind of approaches can hardly handle large-size text.Third is the texture-based method which looks for ″text-like″ background texture areas.These methods apply either wavelet decompositions[21], [22], [23] or Discrete Cosine transform [24], [25] or Gabor filters [26] for feature extraction at the cost of expensive computation involving trainers and classifiers for large databases[27], [28].

Zhuge et.al[29] proposed an effective text detection approach which utilizes gradient amplitude map (GAM) to enhance the edge of an input image, which can overcome the problems of color bleeding and fuzzy boundaries.Background noises are filtered by using two-direction morphological fiers.Maximally stable extremal region (MSER)detction approach is applied to detect text regions with two extreme colors and then the mean

- *P.Sudir is currently pursuing Doctrate degree program in Computer science and engineering in VTU,Belgaum,Karnataka, India,PH-0809844383890. E-mail: sudirhappy@gmail.com*
- *Dr.M.Ravishankar is Serving as Principal,VVIT,Mysore, karnataka, India,PH-0809845550525,E-mail: ravishankamcn@gmail.com*

intensity of the regions as the graph cuts' label set and the Euclidean distance of three channels in HSI color space as the graph cuts smooth term are used to get optimal segmentations.Chucai et al[30]. proposed a framework consisting of 3 main steps: boundary clustering,stroke segmentation and string fragment classification.In boundary clustering,a new bigram color uniformity based method is proposed to model both text and attachment surface.For text discrimination,Pixel of interests are located with maximums of anticompatible Gabor filters and sub-region-based statistical features of oriented histogram,gradient and stroke width are then extracted and classified with an SVM maps of gradient,stroke distribution and stroke width.

Survey of previous works indicate that not much of attention is paid towards multioriented text detection.Variation in lighting and unpredictable transformations makes multioriented text detection more difficult.Extraction of multi-oriented special effects was addressed [31] but limited to unidirectional graphics text.Shivakumar et al.[32] addressed the multi-orientation issue through an algorithm based on Laplacian and skeletonization approach.Higher false positive rate and misdirection rate for non horizontal text were the drawbacks of this method.Yong et al.[33] proposed a corner and skeleton based method for arbitrarily oriented text which is quiet robust only when the corners are detected effectively and fails in case of low resolution Videos.In this paper a multi oriented text detection and extraction method which involves Texture feature extraction from Block wise covariance matrix and MSER based text region verification is proposed.It can simultaneously detect both darker and brighter multi-oriented text without any prior assumptions.In addition, proposed method does not require any prior training tasks,hence it is computationally time efficient when compared to any other machine learning based methods proposed in recent times.

2 Proposed method

2.1 Text Region Candidate Extraction

Text localization problems have been quite successfully handled using Morphological filters. In [34] a scheme utilizing morphological filters based Text Regions extraction method and heuristic based non-text regions filtering were proposed. Assuming High contrast and geometrically similar video text we propose a morphological approach to detect video text.First erode operator is applied (13X50 rectangular Structural element(S)) on the input image, generating the output image f1.

$$f_1 = f\Theta s \tag{1}$$

Extraction of marked objects, Detection of bright regions enclosed by dark pixels, detection or removal of image border touching objects, object holes detection and filling, spurious high or low points filtering are some of the applications where the Morphological reconstruction has been successfully used. Taking input image f as mask and f_1 as the marker, the reconstructed image f_2 is obtained by the following iterative procedure:

1. A structuring element b (7X25 rectangular Structural element) is created
2. Take $h_k = f_2$
3. Repeat:$h_{k+1} = (h_k \oplus b) \cap f_1$, until $h_{k+1} = h_k$
4. $f_2 = h_{k+1}$

Morphological reconstruction using 8-connected neighborhoods of the image f_1 under the image f, generates the output image f_2.Next apply the dilate operator by a 2X1 rectangular Structural Element (S_1) to the image f_2, generating the output image f_3.

$$f_3 = f_2 \Theta S_1 \tag{2}$$

Next subtract original frame from image f_3, generating the output image f_4

$$f_4 = f_3 - f; \tag{3}$$

Image f_5 is derived by taking the minimum value among image f_3 and original image f.Next once again morphological reconstruction is performed using 8-connected neighborhoods of the image f_4 using mask f_5 to generate the output image as shown in figure 1(d).

2.2 Maximally Stable Extremal Regions (MSER) based Text Region Candidate selection

From a given image MSERs (co-variant regions) are extracted using MSER algorithm.[35] MSERs are gray-level sets of the image which have a stable connected component in which all pixels have either higher (bright Extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary.

The MSER extraction involves the following steps:

- Perform a luminance thresholding by Sweeping threshold of intensity from black to white.
- Extract connected components (”Extremal Regions”).
- Find a threshold when the local minimum of the relative growth of Extremal region is ”Maximally Stable”.
- Approximate a region with an ellipse (this step is optional).
- Regions descriptors are kept as features.

Edge enhanced MSER on morphologically detected text regions are extracted as proposed[36].Binary image are obtained and the foreground CCs are considered as text candidates as shown in figure 1(e).In the past,MSER based method have been successful in case of the benchmark data,i.e.,ICDAR 2011 Robust Reading Competition[37] and has produced good results but still MSER algorithm detects a large number of non-characters,most of the character candidates need to be removed before further processing hence we adopt a Covariance Matrix feature extraction based approach for text candidate selection.

2.3 Covariance based Text Region Candidate selection

Gradients are considered as important parameters for image perception and analysis and become key features in various local descriptors in the scope of computer vision and image/video processing. The Gradient image as shown in Figure 1(b) is subdivided into k x×k blocks of size mxm where m is power of two.For each k block, the covariance matrix is computed given by

$$\psi = \frac{1}{m} \sum_{j=1}^{m} \left(\left(k_j - k \right)^2 \left(k_j - k \right) \right) \tag{4}$$

The Standard deviation value of the covariance matrix of each block is computed.The K-means thresholded map of the Standard deviation matrix is shown in figure 1(c).The block whose Kmeans thresholded value is 1 is treated as Text block.The resultant Text Detection output after combining with MSER detected text region is shown in figure1(f).Next heuristic filter methods are adopted in order to reject falsely detected regions like too small,too thin CCs,large CCs based on aspect ratio and size of CCs.Figure 1(g) shows the result after the refinement process.

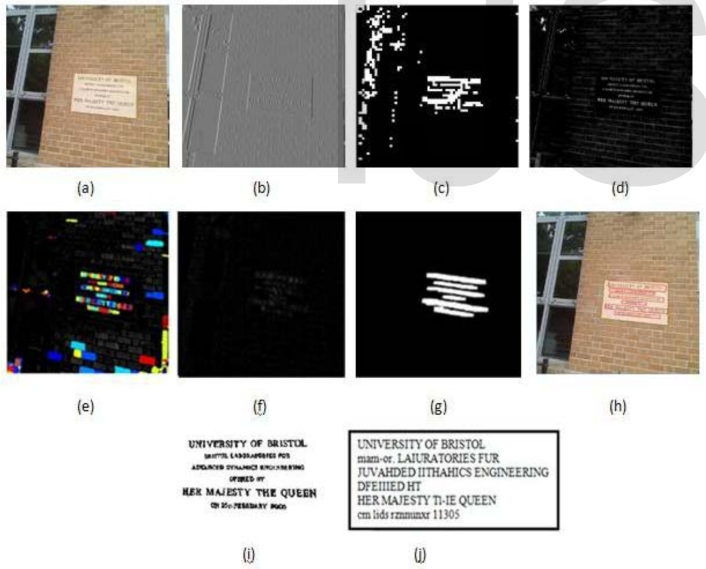


Fig 1 (a) Video frame (b) Gradient image (c) K-map result after Covariance based detection method (d) Morphology output (e) MSER output (f) Covariance based detection and MSER detection combined output (g) Dilation output (h) Detected Text Region (i) Binarized output (j) OCR output

2.4 Text Binarization using Watershed Segmentation and adaptive thresholding

Text Binarization process follows text detection and text localization process in a standard framework of Video OCR.Watershed algorithms[38],[39],[40] are based on the behavior of water that always flows down to lower regions and the watershed divides regions based on the minima that water approaches.This paper proposes the binarization method of videotext CCs based on a water flow

model,in which an extracted CC is divided into two regions,characters and backgrounds,by thresholding the amount of filled water.

The Steps invoved in binarization are

- Supress Noise in each Color channel by applying a median filter of Size 3X3.
- PerformWatershed Segentation on each color channel(Figure 2(b)).The results are shown in Figure 2(c).
- Using the Flooding transform,41weakly significant basins are merged with its neighbors in each color channel.This step is iterated until only strong significant basins are present.
- Choosing Stronger basins from each channel and adaptive threshoding yields the ouput as shown in figure 2(d).
- The OCR output (Figure 2(f))for the binarized segment (Figure 2(e)) is shown

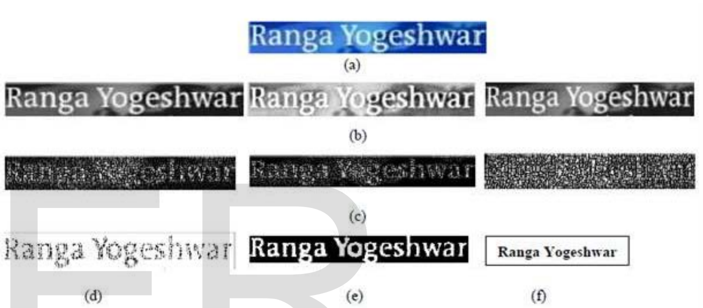


Fig 2 (a) Videotext Segment (b) R,G,B Channel (c) Watershed Segmentation Result on each channel (d) Segmented and thresholded Strong basins (e) Binarized output (f) OCR output

3 Experimental Results

Proposed Method has been tested on two public datasets,First is the ICDAR 2013 Robust Reading Competition Challenge 342 dataset which consists of 28 videos in total: 13 videos for the training set and 15 for the test set.The scenarios in the videos include walking outdoor,shopping in grocery stores, driving and searching for directions within a building.Each video is around 10 seconds to 1 minute long capturing scenes from real-life situations using different types of cameras.The second dataset is the YouTube Video Text (YVT) (http://www.yovisto.com/labs/VideoOCR/) dataset extracted from YouTube.The dataset contains a total of 30 videos of which 15 are from the training set and 15 are from the testing set.Each video has HD 720p quality,30 frames per second and 15-second duration. Following performance criterion has been defined for horizontal and multioriented videotext.Actual Text Blocks (ATB) in the images are manually counted in the dataset to evaluate the accuracy of the text blocks detected.

To judge the correctness of the text blocks detected, we manually count Actual Text Blocks (ATB) in the images in the dataset.

- Truly Detected Block (TDB): A detected block that contains at least one true character. Thus,a TDB may or may not fully enclose a text line.
- Falsely Detected Block (FDB): A detected block that does not contain text.

The performance measures are defined as follows.

a) Recall (R) = TDB / ADB

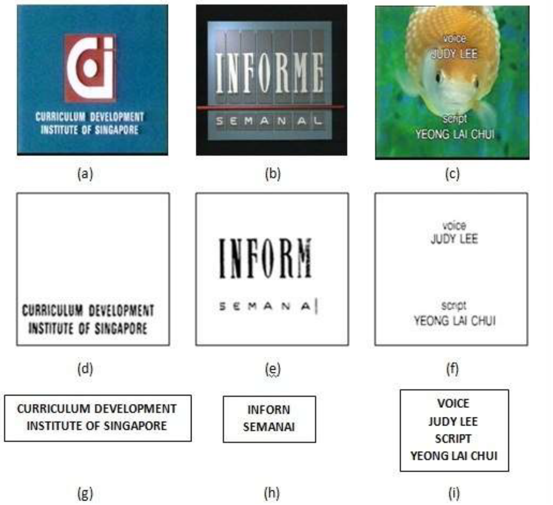
b) Precision (P) = FDB / (TDB +FDB)

c) F-measure (F) = 2PR / (P + R) e)

d) Average Processing Time (APR)

Average Processing Time (APT) per frame is also considered as a measure to evaluate the proposed method. Proposed methods performance has been compared with method[43] which proposes to extract Maximally Stable Extremal Regions (MSERs) as character candidates based on the strategy of minimizing regularized variations.Single-link clustering algorithm clusters Character candidates into text candidates,where a novel self-training distance metric learning algorithm helps in learning distance weights and threshold of the clustering algorithm.The results obtained by Xu method and our proposed methods on frames obtained from ICDAR 2013 video and YVT video dataset are as shown in Figure 5.Experimental results show Xu method has been outperformed by the proposed method as indicated in Table 1 and Table 2.Obtained results show that there is still scope for improvement because of the challenging videos present in the datasets.

The proposed binarization method is tested on 104 Video frame set Collected from German TV news Program and the test set from the Mediaglobe project,which are subsequently referred to as MS testset,TV news testset and MG testset respectively.The proposed binarization method is compared with the well known thresholding based binarization techniques like.[44],[45],[46].The Achieved text recognition accuracy is as shown in table 3 which proves that the proposed method outperforms the results of all other reference procedures.


CURRICULUM DEVELOPMENT INSTITUTE OF SINGAPORE
INFORME
SEMANAL
voice
JUDY LEE
script
YEONG LAI CHUI
(a)
(b)
(c)
CURRICULUM DEVELOPMENT INSTITUTE OF SINGAPORE
INFORM
SEMANA
voice
JUDY LEE
script
YEONG LAI CHUI
(d)
(e)
(f)
CURRICULUM DEVELOPMENT INSTITUTE OF SINGAPORE
INFORN SEMANAI
VOICE JUDY LEE SCRIPT YEONG LAI CHUI
(g)
(h)
(i)

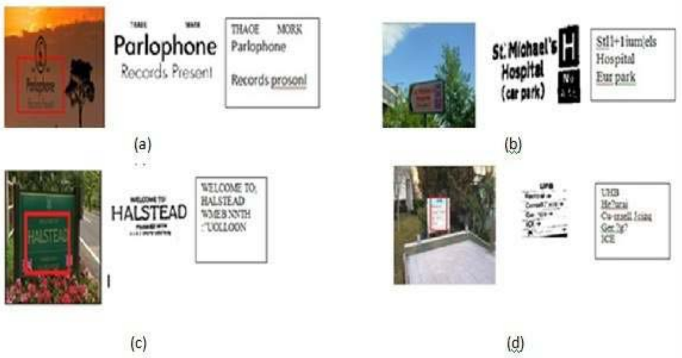


Fig 3 Text Detection and Extraction Recognition examples on the Hua's Dataset

Fig 4 Text Detection and Recognition examples on the ICDAR dataset

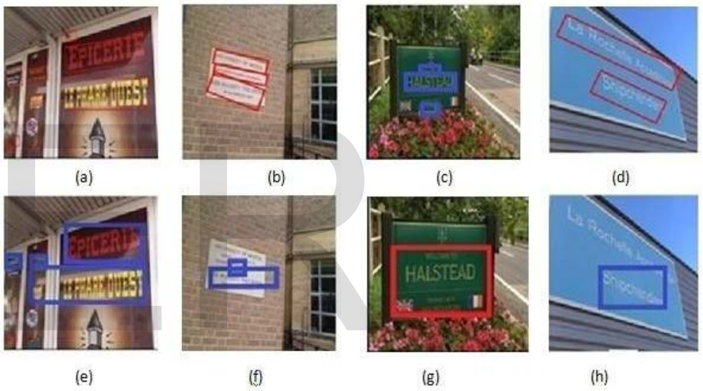


Fig 5 Text detection results (a,b,c,d) Proposed method,(e,f,g,h) Xu method

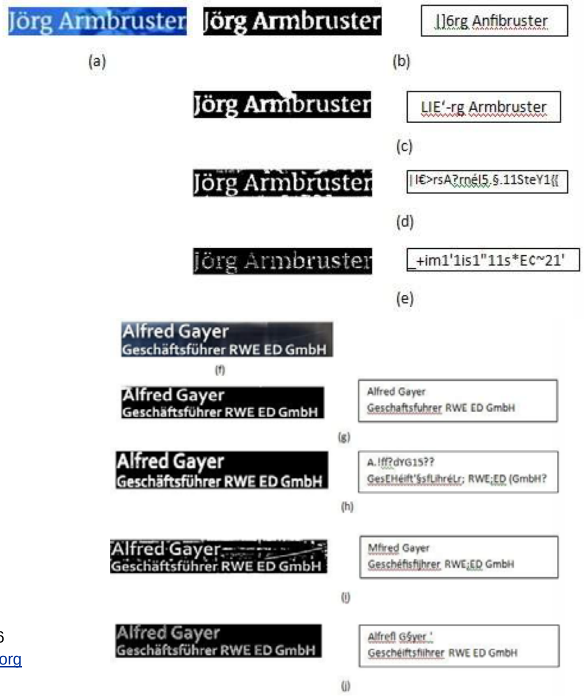


Fig 6 Text Recognition results (a) Proposed method,(b) Otsu method (c) Niblack Method (d) Sauvola Method

Table1 Experimental result of proposed method for YVT dataset

Method	Recall	Precision	F-Measure	APT(Secs
Proposed	0.63	0.85	0.72	0.78
Xu Method	0.62	0.72	0.65	1.2

Table2 Experimental result of proposed method for ICDAR dataset

Method	Recall	Precision	F-Measure	APT(Secs
Proposed	0.67	0.73	0.69	0.83
Xu Method	0.61	0.67	0.61	1.1

Table 3 Performances of different text binarization methods evaluated on the various dataset's

Method	Correct Characters	Correct Words	Char Accuracy	Word Accuracy
Otsu	2087	301	0.6	0.57
Niblack(k=0.5)	1986	287	0.57	0.54
Sauvola(k=0.01)	1400	213	0.4	0.4
Proposed	2413	352	0.72	0.66

4 Conclusion

A Novel Method has been proposed in this paper in which MSER and covariance matrix feature extraction are combined for text detection in video frames.Efficient initial Morphology-based preprocessing stage to detect MSER regions and the Covariance matrix based approach for texture detection and watershed based text binarization forms the main contribution of this work.The experimental results show that the proposed method performs well for both horizontal and nonhorizontal text. The obtained result proves the efficiency of the proposed method.

References

[1] R. G. Jones, "Emerging technologies: Mobile apps for language learning," Language Learning and Technology 15, 2–11 (2011).

[2] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," IEEE Transactions on Image Processing 23(7), 2972–2982 (2014).

[3] A.Jain, X.Peng, X.Zhuang, P.Natarajan, and H.Cao, "Text detection and recognition in natural scenes and consumer videos," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1245–1249(2014).

[4] S. L. Khan, A.Mohammed Salim and S. S. Solanki, "A review on: Video searching using speech and video text information," International Journal of Electrical, Electronics and Computer Engineering 3(2), 169 (2014).

[5] B.Sun, Y.Wu, and C.Chao, "Detecting and extracting text in video via sparse representation based classification method," In Information Science and Applications,Springer, 351–358,(2015).

[6] P. V.Khare and P.Raveendran, "A new histogram oriented moments descriptor for multioriented moving text detection in video," Expert Systems with Applications 42(21), 7627–7640 (2015).

[7] R.Minetto, N.Thome, M.Cord, and J.Stolfi, "Snoopertrack: Text detection and tracking for outdoor videos," 18th IEEE International Conference on Image Processing (ICIP), 351–358 (2011).

[8] H.Jiang, L.Guizhong, Q.Xueming, N.Nan, G.Danping, L.Zhi, and S.Li, "A fast and effective text tracking in compressed video," Tenth IEEE International Symposium on Multimedia, 136 – 1418 (2008).

[9] S. Messelodi and C. M. Modena, "Scene text recognition and tracking to identify athletes in sport videos," Multimedia tools and applications 63(2), 521–545 (2013).

[10] Z.Saidane, C.Garcia, and J.Dugelay, "The image text recognition graph (itrg)," International Conference on Multimedia and Expo, 266–269 (2009).

[11] Phan and T. Quy, "Recognition of video text through temporal integration," 12th International Conference on Document Analysis and Recognition (ICDAR) (2013).

[12] Zhang, Jing, and R. Kasturi, "Extraction of text objects in video documents: Recent progress," In The Eighth IAPR International Workshop on Document Analysis Systems,IEEE, 5–17 (2008).

[13] J.Yi, Y.Peng, and J.Xiao, "Color-based clustering for text detection and extraction in image," In Proceedings of the 15th international conference on Multimedia,ACM, 847–850 (September 2007).

[14] D.Karatzas and A.Antonacopoulos, "Text extraction from web images based on a split-and merge segmentation method using color perception," in Proc. IEEE Int. Conf. Pattern Recognition, 634–637 (2004).

[15] C.Yi and Y.Tian, "Localizing text in scene images by boundary clustering, stroke segmentation and string fragment classification," IEEE Trans. Image Process 21(9), 4256–4268 (Sep.2012).

[16] C. Liu, C. Wang, and R. Dai, "Text detection in images based on unsupervised classification of edge-based features," Proceedings of Eighth International Conference on Document Analysis and Recognition,IEEE (2005).

[17] S. Palaiahnakote, W. Huang, and C. L. Tan, "An efficient edge based technique for text detection in video frames," The Eighth IAPR International Workshop on Document Analysis Systems,IEEE (2008).

[18] A.Mosleh, B.Nizar, and A. Hamza, "Image text detection using a bandlet-based edge detector and stroke width transform," In BMVC (2012).

[19] Phan, T. Quy, P. Shivakumara, and C. L. Tan, "Text detection in natural scenes using gradient vector flow-guided symmetry," 21st International Conference on Pattern Recognition (ICPR), 3296–3299 (2012).

[20] R. Huang, P. Shivakumara, and S. Uchida., "Scene character detection by an edge-ray filter," in Proc. IEEE Int. Conf. Doc. Anal. Recognition, 462–466 (2013).

[21] S. Prakash and M. Ravishankar, "An effective approach towards video text recognition," In Advances in Signal Processing and Intelligent Recognition Systems,Springer International Publishing,323–333 (2014).

[22] P.Shivakumara, A.Dutta, C.L.Tan, and U.Pal, "Multi-oriented scene text detection in video based on wavelet and angle projection boundary growing," Multimedia tools and applications 72(1), 515–539 (2014).

[23] P.Shivakumara, T. Q. Phan, and C. L. Tan, "A robust wavelet transform based technique for video text detection," 10th International Conference on Document Analysis and Recognition,ICDAR'09,IEEE, 1285–1289 (2009).

[24] Q. Xueming, G. Liu, H. Wang, and R. Su, "Text detection, localization, and tracking in compressed video," Signal Processing: Image Communication 22, 752–768 (2007).

[25] Goto and Hideaki, "Redefining the dct-based feature for scene text detection," International Journal of Document Analysis and Recognition (IJDAR) 11(1), 1–8 (2008).

[26] P.Sudir and M.RaviShankar, "A new log gabor approach for text detection from video," International Journal of Signal Processing Systems 2(1), 1–6 (2014).

[27] P.Shivakumara, R.P.Sreedhar, T.Q.Phan, S.Lu, and C.L.Tan, “Multio riented video scene text detection through bayesian classification and boundary growing,” IEEE Transactions on Circuits and Systems for Video Technology 22(8), 1227–1235 (2012).

[28] Wei, Y. Cheng, and C. H. Lin, “A robust video text detection approach using svm,” Expert Systems with Applications 39(12), 10832–10840 (2012).

[29] Y.Z.Zhuge and H.C.Lu, “Robust video text detection with morpholog ical filtering enhanced mser,” Journal of Computer Science and Tech nology 30(2), 353–363 (2015).

[30] C.Yi and Y.Tian, “Localizing text in scene images by boundary clus tering,stroke segmentation and string fragment classification,” IEEE Trans. Image Process 21(9), 4256–4268 ,(Sep.2012).

[31] D.Crandall, S.Antani, and R.Kasturi, “Extraction of special effects caption text events from digital video,” International Journal on Doc ument Analysis and Recognition 5, 138–157 (2003).

[32] P.Shivakumara, T.Q.Phan, and C.L.Tan, “A laplacian approach to multioriented text detection in video,” IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2), 412–419 (2011).

[33] Z.Yong and L.Jianhuang, “Arbitrarily oriented text detection using geodesic distances between corners and skeletons,” 1st International Conference on Pattern Recognition (ICPR), 1896 –1899 (2012).

[34] Y.Hasan and L.J.Karam, “Morphological text extraction from image,” IEEE Transactions on Image Processing , 1978–1983 (2000).

[35] J.Matas, O.Chum, M.Urban, and T.Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” Image and Vision Computing 22, 761–767 (2004).

[36] Huizhong.Chen, Sam.S.Tsai, D. Georg.S, M.Chen, Radek.G, and Bernd.R, “Text detection in natural images with edge-enhanced max imally stable extremal regions,” IEEE International onference on Im age Processing, 2609–2612 (2011).

[37] A.Shahab, F.Shafait, and A.Dengel, “Icdar robust reading competition challenge 2: Reading text in scene image,” International Conference on Document Analysis and Recognition (ICDAR), 1491–1496 (2011).

[38] M.Baccar, L.A.Gee, R.C.Gonzalez, and A.Abidi, “Segmentation of range images via data fusion and morphological watersheds,” Pattern Recognition 29(10), 1673–1687 (1996).

[39] D.Li, G.Zhang, Z. Wu, and L. Yi, “An edge embedded marker-based watershed algorithm for high spatial resolution remote sensing image segmentation,” IEEE Transactions on Image Processing, 19(10), 2781–2787 (2010).

[40] S. Beucher and F. Meyer, “The morphological approach to segmenta tion: the watershed transformation,” In Mathematical Morphology in Image Processing , 433–481 (1993).

[41] M.Couprie and G.Bertrand, “Topological gray-scale watershed trans form,” In Proc. of SPIE Vision Geometry 3168, 136–146 (1997).

[42] D.Karatzas, F.Shafait, S.Uchida, and M.Iwamura, “Robust reading competition,” ICDAR, 1484–1493 (2013).

[43] C.Y.Xu, Y.Xuwang, H.Kaizhu, and Hong, “Robust text detection in natural scene images,” ICDAR, 197–203 (2013).

[44] N.Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Syst., Man, Cybern (1979).

[45] W.Niblack, An Introduction to Digital image processing, Prentice Hill (1986).

[46] J.J.Sauvola and M.Pietikainen, “Adaptive document image binariza tion,” Pattern Recognition (2000).