A Robust Symmetry-based Method for Scene/Video Text Detection Through Neural Network

Yirui Wu[†‡] Wenhai Wang[‡] Shivakumara Palaiahnakote[§] Tong Lu[‡⋆]
[†] College of Computer and Information, Hohai University
[‡]National Key Lab for Novel Software Technology, Nanjing University
[§]Department of Computer System and Information Technology, University of Malaya
{wuyirui@hhu.edu.cn;wangwenhai362@163.com;shiva@um.edu.my;lutong@nju.edu.cn}

***Abstract*—Text detection in video/scene images has gained a significant attention in the field of image processing and document analysis due to the inherent challenges caused by variations in contrast, orientation, background, text type, font type, non-uniform illumination and so on. In this paper, we propose a novel text detection method to explore symmetry property and appearance features of text for improved accuracy and robustness. First, the proposed method explores Extremal Regions (ER) for detecting text candidates in images. Then we propose a novel feature named as Multi-domain Strokes Symmetry Histogram (MSSH) for each text candidate, which describes the inherent symmetry property of stroke pixel pairs in gray, gradient and frequency domains. Furthermore, deep convolutional features are extracted to describe the appearance for each text candidate. We further fuse them by Auto-Encoder network to define a more discriminative text descriptor for classification. Finally, the proposed method constructs text lines based on the classification results. We demonstrate the effectiveness and robustness detection results of our proposed method by testing on four different benchmark databases.**

***Index Terms*—text detection; symmetry property; convolutional network; deep learning; auto-encoder network;**

I. Introduction

Text detection in video/scene images provides important cues for several content-based retrieval applications, such as scene/video analysis and understanding and semantic image/video retrieval [1]. We classify methods for text detection into two categories: sliding window based methods [2], [3] and character region based methods [4]–[7]. Sliding window based methods adopt a window to move over the image and calculate the probability of presence of text on the basis of local image features. However, the computation complexity of such scheme is expensive due to window operation on whole image. Character region based methods explore the inherent properties of characters to group pixels into characters, such as color, intensity or stroke-width. For example, Epshtein et al. [4] consider the constant stroke width - the distance between two parallel edges - to identify the presence of the text in the images. Rather than regions of constant stroke width, Neumann and Matas [6] find Extremal Regions (ERs) which have stable intensity distribution and consider them as candidates for text detection.

Recently, deep neural networks, such as Convolutional Neural Network (CNN) [8]–[10] and Auto-Encoder (AE) network [11], are proposed widely to exploit appearance features to overcome the drawbacks of the conventional methods. However, it is noted that to achieve high performance of theses methods, the dataset for training need to be well designed at large scale. In other words, without a proper framework and suitable training samples, deep learning structure results in overfit and undesirable results for many situations, such as multi-orientation, multi-script or multi-font scene/video text detection. These factors motivate us to exploit geometric features of text to increase the robustness and accuracy for text detection. Inspired by the fact that character exhibits double and parallel edges with uniform distances between the pixels in the parallel edges [4], texts in video and scene images have common intrinsic property of symmetry, which is different from natural objects [12]. This observation inspired us to propose symmetry features in different domains for text detection in video and scene images.

In this paper, we propose a novel method for robust text detection by exploiting the symmetry property of text. By incorporating powerful symmetry property with appearance features, we utilize the intrinsic properties of text to improve robustness of detection for various and ambiguity scenarios. The method achieved good and consistent results on popular datasets for scene and video text detection, which proved that our idea improved the robustness of text detection by combining the strengths of MSSH and CNN.

The main contributions of this paper are two folds: Introducing the Multi-domain Stroke Symmetry Histogram (MSSH) feature based on the inherent symmetry property represented by stroke pairs, which successfully captures the characteristics of text. An highly-efficient and novel way of integrating MSSH feature with deep convolutional network is proposed for text detection in video/scene images. Moreover, the proposed method does not reply on the number of training samples, parameter setting and so on as in deep learning approaches.

This work was supported by the Science Foundation of JiangSu under Grant BK20170892, the Fundamental Research Funds for the Central Universities under Grant 2013/B16020141, the open Project of the National Key Lab for Novel Software Technology in NJU under Grant KFKT2017B05, the Natural Science Foundation of China under Grant 61672273, Grant 61272218, Grant 61321491, and the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant BK20160021.

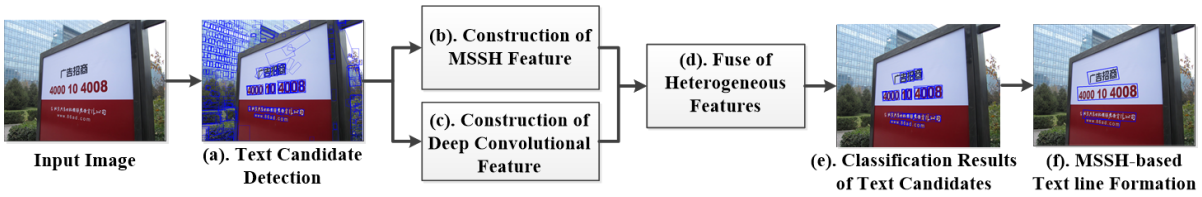

Fig. 1. The framework of the proposed method: (a) Text candidates detection, (b) the construction of MSSH feature, (c) the construction of deep convolutional feature, (d) the fuse of MSSH and convolutional features, (e) classification results of text candidates and (f) MSSH-based text line formation.

II. The Proposed Method

In this section, we propose a novel method to explore symmetry property and appearance feature of text for robust text detection. Fig. 1 shows the overview of the proposed method, which consists of the following steps: (a) text candidates given by detecting and filtering ERs, (b) MSSH feature to extract symmetry property of text candidate in multi-domains, (c) deep convolutional feature to extract the appearance feature of text candidates, (d) the fuse of MSSH and deep convolutional features with the help of Auto-Encode network to extract distinct features for classification, (e) classification of text and non-text candidates, and (f) the usage of MSSH feature to further group the text candidates into text lines.

A. Text Candidate Detection

As noted from literature that ER helps in extracting dominant regions in an image, we explore ER concept to detect dominant regions which are named as text candidates. Most existing methods adopt MSER concept for text candidates detection, which is not robust to multiple color, distortion and low-contrast texts. Fig. 2 shows several text candidate detection results for images affected by severely non-uniform illumination and non-connected components (Chinese Characters) using MSER [13] and the proposed ER method. It is observed from Fig. 2 that MSER fails to detect character as one text candidate, while the proposed ER method could detects characters as text candidates for both images. Specifically, we propose to generate text candidate, say $\{e_i\}$ of the input image I by detecting ERs in multi-channels $\{C_l|l = 1...6\}$, i.e. RGB and HSV.

Due to complex background and contrast variations, the conventional ER method alone may not be sufficient to detect text candidates accurately. Therefore, we propose the following filters to reduce non-text candidates generated by ER method.

1) Filter based on geometric features: Since characters usually have similar geometric appearance, we estimate the ratio r_1 between the area and diameter of ER, the ratio r_2 between the width and height of ER and the holes r_3 of ER which is represented by Euler number after binarization.

2) Filter based on Intensity Distribution: Inspired by the fact that character have uninform color values, the proposed method discards the text candidates which have high variation in intensity values. Supposed that ER should contain text and non-text parts, we first perform histogram operation on intensity values and then adopt mean of the maximal and second-maximal value of the histogram as the split value. Based on the split value, we calculate the variance of intensity distribution D_i of ER by the following equation:

$$D_i = \frac{n_t \cdot \sum\limits_{x \in e_{i,t}} (I_x - M_{e_{i,t}})^2 + n_b \cdot \sum\limits_{x \in e_{i,b}} (I_x - M_{e_{i,b}})^2}{n_t + n_b} \tag{1}$$

where $e_{i,t}$ and $e_{i,b}$ represent the text and non-text of e_i respectively, n refers to the number of pixels and M represents the mean value. Essentially, the goal of filtering with r_1, r_2 and D_i is to adaptively delete outliers of ERs. Note we simply discard ERs with no holes for filtering of r_3. Any ER which passes all these filters would be regarded as text candidates. After filtering, we could get a convinced and small set of text candidates E for subsequent steps (the number of E for one input image is usually up to 150).

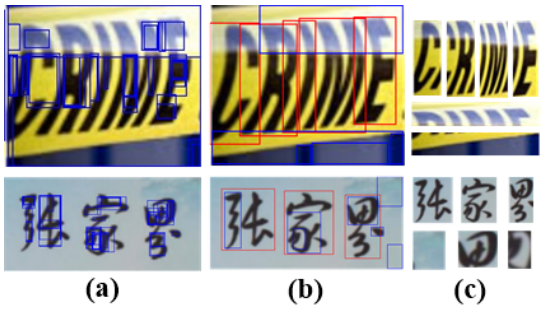

Fig. 2. The text candidates detection results by MSER and our method are shown in (a) and (b) respectively, while (c) represents the samples of text candidates generated by our method. Note that red rectangles in (b) indicates the candidates regions contains texts.

B. Construction of MSSH Feature

This subsection presents MSSH feature extraction from text candidates given by the previous subsection, which describes the inherent symmetry property for each text candidate.

Text has regular pattern, such as constant stroke width and uninform color values [4]. Besides, the intensity values of each stroke pixel pair, which are on the parallel strokes, tend to be similar due to homogenous backgrounds [14], [15]. The proposed method thus describes these unique observations by designing descriptor named as Multi-domain Stroke Symmetry Histogram (MSSH), to integrate these two kinds of symmetry property, i.e. intra-class and inter-class symmetry. Specifically,

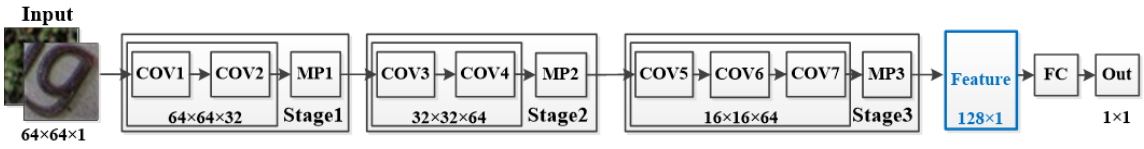

Fig. 3. An illustration of the architecture of the proposed CNN model. The convolutional, max-pooling and full-connect layers are represented as COV, MP and FC respectively, while the proposed deep convolutional feature is represented by blue square.

intra-class symmetry means values of each stroke pixel pair tend to be similar in intensity and gradient due to homogenous backgrounds, while inter-class symmetry means the stroke sequences in one text candidate are inclined to be close in length, intensity value distribution and low-frequency pattern. In this way, MSSH feature which comprises symmetry properties of text helps distinguish between text and non-text candidates. Note that we extract the symmetry properties from different domains, i.e. color spaces, gradient domain and frequency domain, which strengthens the discriminative power of the proposed MSSH feature.

To construct MSSH feature, we first perform histogram operation over symmetry features with predefined quantization (bins) and then concatenate all the histograms as one descriptor. Supposing that $\{p, q\}$ is one sample stroke pixel pair of ith text candidate e_i, we could compute the symmetry feature F_j corresponding to the jth symmetry property as follows:

$$F_j(e_i) = \begin{cases} f_h(|I_j(p) - I_j(q)|) & \text{if } j \in \{V, G_m\} \\ f_h(cos\langle I_j(p), I_j(q)\rangle) & \text{if } j = G_o \\ f_h(f_\xi(s, j)) & \text{if } j \in \{Sw, Md, Pa\} \end{cases} \tag{2}$$

where function $f_h()$ represents the histogram operation, s means the sequence between p and q, $\langle\rangle$ means the angle between two vectors, $\{V, G_m, G_o\}$ represent the intra-class symmetry properties of intensity, gradient magnitude and gradient orientation respectively, $\{Sw, Md, Pa\}$ represent the inter-class symmetry properties of stroke width, value distribution and low frequency pattern respectively, and function $f_\xi()$ assigns the score values for inter-class symmetry properties based on stroke sequences and could be defined as:

$$f_\xi(s, j) = \begin{cases} \parallel s \parallel & \text{if } j = Sw \\ D_s / M_s^2 & \text{if } j = Md \\ \sum_{k=0}^{n_l} \omega_k \cdot f_\varphi(s, k) & \text{if } j = Pa \end{cases} \tag{3}$$

where $\|\|$ refers to Euclidean distance. Inspired by the score of stroke width cue in [16], we utilize D_s and M_s, the variance and mean value of intensity values of sequence s, to score the proposed value distribution property Md. Regarding stroke sequence as a signal, wavelet transform help us to transform sequence into frequency domain with different scales. We thus use Haar wavelet transform, represented as function $f_\varphi()$, to extract low-frequency pattern of sequence with different scales varying from 0 to n_l, where n_l represents the total level number (We set $n_l = 3$ by experiments). In other words, we will discard high-frequency coefficients part for levels from 0 to n_l during transform. After transforming, we will score the low-frequency pattern of the sequence by predefined weight vector ω_k, which assigns larger weight for higher level pattern.

To make the resulting feature be the same in size for text candidates with different size, we need to normalize the histogram. The max/min values in histogram are certain for intra-class properties, i.e. $\{V, G_m, G_o\}$, while we rescale values for inter-class properties, i.e. $\{Sw, Md, Pa\}$ to the range from 0 to 1. The reason for rescaling is that we only cares about the distribution of values of inter-class properties, not the detailed values. We thus normalize features F_j of different symmetry properties and concatenate them to be the final MSSH feature by $F_m(e_i) = [F_j(e_i)|j = 1, ..., 6]$ with size $30 \times 6 = 180$.

In fact, conventional Stroke Width Transform often fails to offer accuracy and robust results for low resolution Scene/Video images with complex background, since it could identify wrong pixel pairs as strokes. The symmetry property of stroke pixel pairs we introduced help remove such wrong pairs to improve the text detection performance. After constructing MSSH features, we find text regions share a similar distribution in symmetry values, while symmetry distribution of non-text regions didn't resemble text regions. Therefore, the symmetry distribution helps in distinguishing between text and non-text. Besides, MSSH feature is invariant to rotation, scaling and some extent to distortion.

C. Construction of Deep Convolutional Feature

This subsection presents how to learn the deep convolutional features for text detection by our proposed CNN model.

CNN has gained attention of researchers as it has ability to represent complex data in simple way. This observation leads to learn the highly nonlinear feature representation of appearance from CNN model to discriminate between text and non-text. In other words, we aim to extract features from the proposed CNN, which is trained as a binary classifier to classify given text candidate as text candidate or non-text candidate. Specifically, we modified the structure of currently popular VGG-16 architecture [17] to construct the proposed deep convolutional feature, which is shown in Fig. 3. The input of our CNN is a 3-channel text proposal of size 64×64. Note that we normalize the sizes of all the text candidate to 64×64 to fit with the structure. The first and second convolution layer (COV1,COV2) adopts 32 kernels of size $3 \times 3 \times 1$ with a padding of 1 pixel. After convolving with multiple filter masks, we apply Rectified Linear units (ReLU) as non-linear activation function. This results in $64 \times 64 \times 32$ feature dimension matrix and it is passed to max-pooling layer

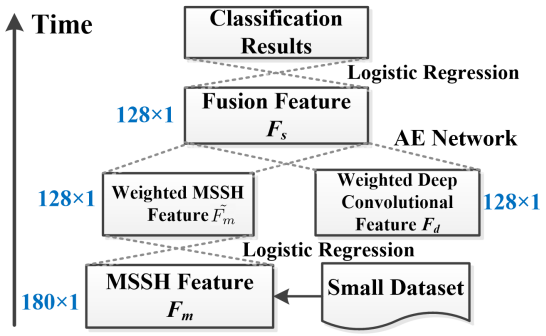

Fig. 4. The fusion model is constructed by Auto-Encoder Network, where blue numbers visualize the dimensions of each feature.

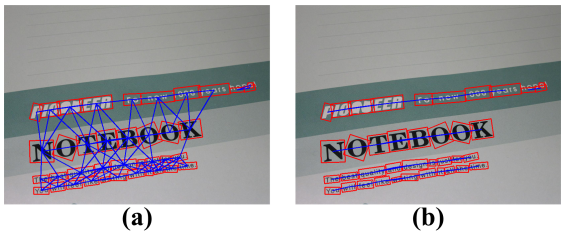

Fig. 5. The result of MSSH-based text line formation. (a) parallel text lines are easy to be confused as one single text line; (b) the proposed method could separate the parallel text lines.

(MP1), which fuses the 2×2 spatial neighborhoods with a stride of 2 pixels. The layers of COV1, COV2 and MP1 form the Stage 1 of the CNN architecture. Except for the number of filter kernels, the parameter configurations for other convolutional and maxpooling layers are the same. In fact, the convolutional layers of our CNN are the core which provide various and hierarchy feature maps of appearance, while the max-pooling layers offer the activation features with the ability of robustness to slight shifts in appearance. The output of the third max-pooling layer is the proposed 128 dimensions deep convolutional feature for text detection, which is then feed to the fully-connected layer (FC) to assign the binary label, i.e. text or non-text for the input text candidates. Essentially, the structure of fully-connecting could globally build classifications with stronger capabilities by containing distributed representations of feature maps in distant parts of the input text proposals.

By simplifying the VGG-16 architecture and designing deep features of 128 dimensions, we achieve great reduction in huge computation burden given by CNN. At the same time, the proposed architecture ensures the capability of extracting of effective appearance features F_d for subsequent steps.

D. Fusion of Heterogeneous Features

The extracted MSSH and CNN features represent different characteristics of characters components. In order to take advantage of both features, we propose to fuse them for classification of text and non-text candidates. The fusing strategy is essential since text detection problem is complex. By fusing heterogeneous features for classification, the proposed method could be robust to multiple text detection challenges caused by diverse factors.

Inspired by work [18] which fuses multimodal data, i.e. audio and video, to learn a shared representation, we adopt Auto-Encoder network to fuse our proposed MSSH feature F_m and deep Convolutional feature F_d, which is shown in Fig. 4. Different from [18] which uses the same stacked RBMs/DBN architecture to represent features before the fusion, we propose CNN and MSSH features instead. The joint training is included to adjust the parameters to be able handle the heterogeneity and produce a more reliable estimate from the heterogeneous data. To speed up the process of fusing, we define pre-fused weights directly as initializations for fusing, since former steps and the fusing step are both designed to classify text or non-text components. Therefore we directly initialize the weights ω_d of the layers from the previously trained CNN for deep convolutional feature F_d. For hand-crafted feature F_m, we adopt a logistic regression (LR) model to assign prefused weights ω_m and reduce dimensions. The proposed LR model will be trained with a small dataset D consisted by labeled text candidates. In fact, the idea of adopting LR for binary classification transforms the training of hand-crafted feature to one fully-connected layer. Such operation is similar to the spirit of FC layer of CNN. The whole process of generating fusion feature F_s could defined as follows:

$$\begin{cases} \{\tilde{F_m}(e_i), \omega_m\} = f_\tau(F_m(e_i); D) \\ F_s(e_i) = f_\mu(\omega_d, F_d(e_i), \omega_m, \tilde{F_m}(e_i)) \end{cases} \tag{4}$$

where function $f_\tau()$ and $f_\mu()$ represents the LR and AE network and $\tilde{F_m}$ refers to the MSSH feature after dimensionality reduction. Note that we keep $\tilde{F_m}$ and F_d to be same in dimensions for equal representations. The training of AE network ends when the validation error rate stops decreasing. During experiments, we find our fusion model could end in less than 10 epochs, which proves the efficiency of our fusing model by adopting pre-fused weights. After fusing, we apply the fusing feature F_s in a random forest model to classify text from set of text candidates.

E. MSSH-based Text Line Formation

This subsection presents text line formation based on classification results of text candidates offered by previous step.

Classified text candidates provide coarse locations of text. To draw bounding box of text line, the proposed method groups the text candidates which are near in distance and share the common properties such as scale, orientation and symmetry. The advantage of such text line formation strategy is that it could extract text line with arbitrary orientations as shown in Fig. 5. Specifically, the proposed method groups multi-oriented text candidates p and q based on the following conditions:

$$\begin{cases} 2/3 \leq |H_p/H_q| \leq 3/2 \\ |\theta_p - \theta_q| \leq \pi/8 \\ |f_d(p,q) - (W_p + W_q)/2| \leq H_p + H_q \\ \|F_m(p) - F_m(q)\| \leq 7 \end{cases} \tag{5}$$

TABLE I
PERFORMANCE OF TEXT DETECTION ON MSRA

Method	Precision	Recall	F-measure
Proposed	0.75	**0.68**	**0.72**
MSSH	0.68	0.60	0.64
CNN	0.74	0.64	0.68
He et al. [8]	**0.81**	0.63	0.71
Yin et al. [19]	0.71	0.61	0.65
Li et al. [16]	0.67	0.61	0.64

TABLE II
PERFORMANCE OF TEXT DETECTION ON ICDAR 2015 SCENE

Method	Precision	Recall	F-measure
Proposed	**0.72**	0.39	**0.50**
MSSH	**0.72**	0.37	0.49
CNN	0.70	0.32	0.44
StradVision-1	0.53	0.46	0.50
NJU Text	0.70	0.36	0.47
AJOU	0.47	**0.47**	0.47

TABLE III
PERFORMANCE OF TEXT DETECTION ON SVT

Method	Precision	Recall	F-measure
Proposed	**0.78**	**0.66**	**0.72**
MSSH	0.69	0.54	0.61
CNN	0.77	0.57	0.66
Mosleh [21]	0.76	**0.66**	0.71
Yin et al. [19]	0.41	**0.66**	0.51
Li et al. [16]	0.74	0.60	0.66

TABLE IV
PERFORMANCE OF TEXT DETECTION ON ICDAR 2013 VIDEO

Method	Precision	Recall	F-measure
Proposed	**0.78**	0.67	**0.72**
MSSH	0.75	0.53	0.62
CNN	0.75	0.56	0.64
Li et al. [16]	0.46	**0.70**	0.56
Yin et al. [19]	0.64	0.57	0.60
Wu et al. [22]	0.63	0.68	0.65

where H, W and θ represent the height, width and orientation of text region respectively, and $f_d(p, q)$ refers to the distance between center of p and q. Note that all the parameters in Eq. 5 are determined experimentally. If two text candidate satisfy all these conditions, the proposed method considers these text candidates for grouping.

After grouping text candidates into text lines, we find that some parallel text lines are easy to be confused as one single text line shown in Fig. 5(a). We thus propose to revise the confused text line based on the orientations of connection lines. Specifically, we keep one connection line t as part of text lines only if its orientation t_θ is equal to the orientation value appearing most frequently in the image. Such process could be defined as follows:

$$\tilde{t} = \{t, if\ t_\theta = \max_C f_h(\{t_\theta\})\} \tag{6}$$

where function $f_h()$ represents the histogram function, C refers to the orientation value which appears most frequently in the histogram. The revision result of text lines is shown in Fig. 5(b), where we can notice the parallel text lines are separated well.

III. EXPERIMENTS

To evaluate the proposed method, we consider four benchmark databases, namely, MSRA, ICDAR 2015 scene, SVT and ICDAR 2013 video. Note that MSRA, ICDAR 2015 scene, SVT are used for validating text detection in natural scene images, while ICDAR 2013 video is used for validating text detection in video frames. To measure results of text detection in natural scene images and video frames, we follow standard evaluation scheme as in the ICDAR robust competition [20], which adopts Recall, Precision and F-measure to evaluate the performance of text detection.

In this work, two models need to be trained with training sets: MSSH-realted logistic regression model and CNN. For the purpose of training, we first use subset of training samples from ICDAR 2003 video and MSRA datasets as training set for MSSH-realted logistic regression model. We then change orientations and add gaussian noise to create more training samples for CNN model, which results in 2.8×10^5 training samples for CNN. We train the random forest classifier with the parameters settled by experiments, where the number of the trees is 10, the function to measure the quality of a split is Gini impurity and the maximum depth of the trees is 6.

Table. I - IV gives the detailed statics of the proposed and existing methods for the MSRA, ICDAR 2015 scene, SVT and ICDAR 2013 video datasets, respectively. Since we use standard dataset and evaluation scheme, we report the same results given by the existing methods as in the paper. It is observed from Table. I - IV that MSSH alone and CNN alone give reasonable results compared to the results of existing methods. MSSH alone is not sufficient to handle the issues of different database, such as video data which suffer from low resolution, complex background and multiple text types, MSRA data which includes arbitrary oriented texts, SVT which suffer from severe complex background and perspective distortion and natural scene data which suffer from different fonts, font size etc. In conclusion, MSSH achieves higher F-score than deep feature if the appearance and shape of characters are obvious as shown in experiment statics on ICDAR 2015 scene. MSSH achieves the nearly same F-score as deep features in ICDAR 2013 video, which proves MSSH is robust to motion effect. If the quality of characters is affected by low resolution and arbitrary rotation, MSSH could be less effective as proved by experiment results in SVT and MSRA. Therefore, to achieve better results for all three type databases, we propose to combine the strengths of MSSH and Deep convolutional networks. It is evident from Table. I - IV that the proposed method achieves the best results in terms of recall or precision or F-measure for all four types databases. This is novelty of the proposed work. This shows MSSH and CNN both contributes to achieve the best results by solving complex issues which remain as open issues for text detection in natural scene and video frames. Therefore, we can assert that the proposed method is robust to natural scene text detection and video text detection unlike existing methods give good results

(a) (b) (c) (d) (e) (f)

Fig. 6. Detection examples of the proposed method on MSRA (a), ICDAR 2015 scene (b,c), SVT (d) and ICDAR 2013 video (e,f).

for one specific database. This is the main advantage of the proposed method compared to the state of the art methods.

Compared with other methods, the consistent top performance achieved on the four datasets demonstrates the effectiveness and generality of the proposed method. By incorporating symmetry information of text, our method even outperforms several full CNN method in f-measure. For example, the f-measure on MSRA by proposed method is 0.72 compared with 0.71 achieved by He et al. [8]. This proves the effectiveness of incorporating geometry information of text. In fact, text have a strong inherent property in geometry, which could be utilized to improve robustness and accuracy. Note that we perform experiments with the proposed and comparative methods on a laptop (2.9GHz 2-core CPU, 8G RAM, Nvidia GeForce GTX 960M and Windows 64-bit OS) and the average time to process with an image is 5.3s, which is much smaller than 16.7s per image achieved by the fully-CNN work, i.e. He et al. [8]. The main reason for shorter processing time lies in the simplification version of VGG structure and effective MSSH extraction processing. Sample qualitative results of the proposed method are shown in Fig. 6 where we can see the proposed method detect text in different complex images well. For instance, our method successfully detects the text of perspective distortion as shown in (e), text of broken strokes, multi-script, multi-orientation has been detected by the method accurately as shown in (a) and (b) text affected by non-uniform illumination is detected successfully as shown in (c) text in complex background is also detected successfully by the method as in (d) and (f).

IV. Conclusions

In this paper, we propose a robust text detection method for scene and video images by fusing symmetry property and appearance features. Initially, the proposed method explores ER to detect text candidates from the input images. Then, the proposed method introduces a novel descriptor called MSSH to verify the text candidates by extracting intra and inter symmetry features of character components. For the same text candidate, the proposed method extracts appearance features using CNN. The strengths of MSSH and CNN features are fused with AE network to achieve efficient and accurate classification results. Furthermore, the MSSH feature are used to group text candidates into text line. Experimental results demonstrate that the proposed method outperforms the existing methods in terms of effectiveness and robustness.

References

[1] Q. Ye and D. S. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, 2015.

[2] L. G. i Bigorda and D. Karatzas, "Multi-script text extraction from natural scenes," in *Proc. ICDAR*, 2013, pp. 467–471.

[3] L. Kang, Y. Li, and D. S. Doermann, "Orientation robust text line detection in natural images," in *Proc. CVPR*, 2014, pp. 4034–4041.

[4] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. CVPR*, 2010, pp. 2963–2970.

[5] L. Neumann and J. Matas, "Text localization in real-world images using efficiently pruned exhaustive search," in *Proc. ICDAR*, 2011, pp. 687–691.

[6] L. Neumann and J. Matas, "Real-time lexicon-free scene text localization and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1872–1885, 2016.

[7] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. CVPR*, 2012, pp. 1083–1090.

[8] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE Trans. Image Processing*, vol. 25, no. 6, pp. 2529–2541, 2016.

[9] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. ECCV*, 2016, pp. 56–72.

[10] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. CVPR*, 2016, pp. 4159–4167.

[11] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proc. NIPS*, 2011, pp. 801–809.

[12] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based text line detection in natural scenes," in *Proc. CVPR*, 2015, pp. 2558–2567.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. BMVC*, 2002, pp. 1–10.

[14] J. Zhang and R. Kasturi, "A novel text detection system based on character and link energies," *IEEE Trans. Image Processing*, vol. 23, no. 9, pp. 4187–4198, 2014.

[15] Y. Wu, P. Shivakumara, T. Lu, C. L. Tan, M. Blumenstein, and G. H Kumar, "Contour restoration of text components for recognition in video/scene images," *IEEE Trans. Image Processing*, vol. 25, no. 12, pp. 5622–5634, 2016.

[16] Y. Li, W. Jia, C. Shen, and A. V. D. Hengel, "Characterness: An indicator of text in the wild," *IEEE Trans. Image Processing*, vol. 23, no. 4, pp. 1666–1677, 2014.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 689–696.

[19] X. Yin, X. Yin, K. Huang, and H. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, 2014.

[20] D. Karatzas et al., "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, 2013, pp. 1484–1493.

[21] A. Mosleh, N. Bouguila, and A. B. Hamza, "Automatic inpainting scheme for video text detection and removal," *IEEE Trans. Image Processing*, vol. 22, no. 11, pp. 4460–4472, 2013.

[22] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1137–1152, 2015.