IEICE TRANS. INF. & SYST., VOL.E107–D, NO.4 APRIL 2024

LETTER

VTD-FCENet: A Real-Time HD Video Text Detection with Scale-Aware Fourier Contour Embedding

Wocheng XIAO[†,††], ***Nonmember***, **Lingyu LIANG**[†,††a)], ***Member***, **Jianyong CHEN**[†,††], ***Nonmember***, ***and*** **Tao WANG**[†††,††††b)], ***Member***

SUMMARY Video text detection (VTD) aims to localize text instances in videos, which has wide applications for downstream tasks. To deal with the variances of different scenes and text instances, multiple models and feature fusion strategies were typically integrated in existing VTD methods. A VTD method consisting of sophisticated components can efficiently improve detection accuracy, but may suffer from a limitation for real-time applications. This paper aims to achieve real-time VTD with an adaptive lightweight end-to-end framework. Different from previous methods that represent text in a spatial domain, we model text instances in the Fourier domain. Specifically, we propose a scale-aware Fourier Contour Embedding method, which not only models arbitrary shaped text contours of videos as compact signatures, but also adaptively select proper scales for features in a backbone in the training stage. Then, we construct VTD-FCENet to achieve real-time VTD, which encodes temporal correlations of adjacent frames with scale-aware FCE in a lightweight and adaptive manner. Quantitative evaluations were conducted on ICDAR2013 Video, Minetto and YVT benchmark datasets, and the results show that our VTD-FCENet not only obtains the state-of-the-arts or competitive detection accuracy, but also allows real-time text detection on HD videos simultaneously.

key words: *video text detection, video, scene text detection*

1. Introduction

Video text detection aims to localize and track text instances in videos. Since most video contains text, text detection is a significant stage in many applications, like video retrieval [1], [3] and autonomous driving [4].

Existing video text detection (VTD) methods can be roughly divided into two categories. One line of works formulate the VTD problem as a special object detection problem. Many of these methods are based on a bottom-up strategy, which modify an object detection or instance segmentation framework to locate components of text instances and then aggregate these components to obtain final outputs [5]. Other lines of works uses point sequences of closed-form curves or bounding boxes with appearance and geometry feature to model the boundary of text instances, and formulates the VTD problem based on a top-down strategy. These methods utilize a tracking framework [6]–[8] with feature fusion [9], [10] to address the variances of motion blur or lighting changes.

It can be found that most of previous VTD methods only focus on improving detection accuracy, but few of them consider the speed issue. Since real-time VTD is significant for many applications, this paper aims to explore the challenging problem of real-time text detection on HD video, whose detection speed is over 30fps (frames per second) on ordinary videos.

According on our preliminary analysis and experiments, directly modifying previous VTD methods to achieve real-time VTD tasks may fail to perform well. The challenges are two-folds. Firstly, most previous methods are trained on video data with reduced resolution, whose performances may be difficult to maintain on HD video. Secondly, the methods based on object segmentation framework or that based on tracking framework may require deliberately-designed network architecture or feature fusion components, which may cause intrinsically high computational complexity and fail to achieve a real-time 30fps speed, as shown in Table 1.

In this paper, we propose a real-time HD video text detection method considering both the issues of accuracy and speed. Based on our preliminary works for text detection on images [2], we use the Fourier Contour Embedding (FCE) signatures to represent arbitrary shaped text contours in the Fourier domain. Then, we propose the scale-aware VTD-FCE method, which adaptively selects the scale of the FCE feature backbone network that is mostly matching to the scale of video text instances in the training stage.

Equipped with the VTD-FCE method, we constructed the VTD-FCENet for real-time video text detection, which has an adaptive lightweight end-to-end architecture to achieve a good balance between detection accuracy and speed. VTD-FCENet consists of a ResNet50 network, a feature pyramid network, three scale-aware prediction heads, and a GPU accelerated post-processing module. Each prediction head contains three branches: a classification branch, a regression branch, and a modeling point adaptation branch. The inter-frame fusion mechanism is introduced to obtain temporal correlation between the preceding and following frames. The first branch predicts the possible text regions and text center regions, the second branch predicts the Fourier

Manuscript received May 18, 2023.
Manuscript revised September 27, 2023.
Manuscript publicized December 7, 2023.

[†]The authors are with South China University of Technology, China.
[††]The authors are also with Guangdong Artificial Intelligence and Digital Economy Laboratory (Pazhou Lab Guangzhou), China.
[†††]The author is with Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), China.
[††††]The author is also with The Key Laboratory of Cognitive Computing and Intelligent Information Processing of Fujian Education Institutions, Wuyi University, China.
a) E-mail: lianglysky@gmail.com
b) E-mail: twang@mju.edu.cn
DOI: 10.1587/transinf.2023EDL8030

Table 1 Comparisons on ICDAR2013 Video, YVT and Minetto Datasets, where detection accuracy is measured by Precision (P), Recall (R), and F-measure (F) and inference speed is measured by frames per second (fps).

Methods	Paper	ICDAR2013 Video				YVT				Minetto			
		P(%)	R(%)	F(%)	S(fps)	P(%)	R(%)	F(%)	S(fps)	P(%)	R(%)	F(%)	S(fps)
Shivakumara et al. [5]	PR'17	61.0	57.0	59.0	-	79.0	73.0	77.0	-	-	-	-	-
Wang et al. [16]	ICMR '18	58.3	51.7	54.4	-	-	-	-	-	88.8	87.5	88.1	-
Wang et al. [17]	PCM '18	71.9	58.7	62.6	-	-	-	-	-	83.0	84.2	83.3	-
Wang et al. [10]	MM '19	69.4	55.1	59.0	-	73.0	67.5	69.2	-	-	-	-	-
Chen et al. [9]	ICME '21	74.6	**64.4**	67.3	11.5	78.7	**78.0**	77.2	12.9	-	-	-	-
Yu et al. [7]	PR '21	81.2	57.5	67.3	11.9	**89.1**	71.0	**79.1**	11.9	92.2	90.6	91.4	11.9
Gao et al. [6]	TIP '21	82.4	59.2	68.9	<1	-	-	-	-	-	-	-	-
FCENet [2]	CVPR '21	**83.8**	54.8	66.2	21.8	81.9	64.3	72.0	30.6	86.1	86.7	86.4	24.0
VTD-FCENet	**Ours**	76.7	63.3	**69.4**	**39.2**	79.4	75.2	77.3	**54.9**	**92.3**	**90.7**	**91.5**	**42.9**

vectors containing text contour information, and the third branch predicts the modeling point number used for post-processing. Finally, the post-processing module reconstructs and aggregates the predicted Fourier vectors and removes redundancies via non-maximum suppression (NMS). VTD-FCENet can be efficiently accelerated via GPU, but it is worth to note that even without GPU acceleration, our VTD-FCENet can achieve real-time detection with good accuracy.

The experimental results have verified the effectiveness and real-time performance of our VTD-FCENet in video text detection. Our method has achieved state-of-the-art performance on the ICDAR 2013 Video [11] and Minetto [12] datasets, and competitive performance on the YVT [13] dataset. Meanwhile, our inference speed is much faster than previous methods, and we can achieve real-time detection in HD input videos.

The main contributions are summarized as follows:

- VTD-FCE method, which models arbitrary-shaped text contours as compact signatures in Fourier domain, is proposed. It adaptively selects the feature scale corresponding to the training text instances and obtain temporal correlations between adjacent frames via frame-level fusion mechanism.
- Based on VTD-FCE, VTD-FCENet is constructed to achieve a real-time video text detection with a lightweight end-to-end architecture. VTD-FCENet can greatly improve its inference speed by GPU acceleration and network optimization while obtaining good detection accuracy.
- Experimental results and comparisons with related methods on ICDAR 2013 Video, Minetto and YVT benchmark datasets show that our VTD-FCENet not only obtains state-of-the-art or competitive on detection accuracy, but also obtains the highest inference speed and achieves real-time text detection on HD videos.

2. Proposed Method

2.1 Scale-Aware VTD-FCE Method

Based on preliminary works for text detection on images [2], which represent arbitrary shaped text contours using Fourier Contour Embedding (FCE) signatures in the Fourier domain, we propose VTD-FCE method with scale-aware and inter frame fusion mechanisms to achieve real-time HD video text detection.

In VTD-FCE, input video stream with s frames can be represented as $\mathbf{V_s} = [F_1, \ldots, F_s]$. Each frame F in stream contains corresponding contours $\mathbf{C}$, which can be represented in the following format:

$$\mathbf{C} = \mathbf{X} + i\mathbf{Y} \tag{1}$$

$\mathbf{C} = [C_1, \ldots, C_m]$ denotes m contours in this frame. $\mathbf{X} = [x_1(t), \ldots, x_m(t)]$ and $\mathbf{Y} = [y_1(t), \ldots, y_m(t)]$ denote spatial coordinates in contours. Note that contour $C(t) = C(t+1)$, $t \in [0, 1]$. We adpot Inverse Fourier Transformation (IFT) to formulate $\mathbf{C}$

$$\mathbf{C} = \sum_{k=-\infty}^{+\infty} \hat{\mathbf{a}_k} e^{2\pi i k} \tag{2}$$

$k \in Z$ denotes frequency, $\hat{\mathbf{a}_k} = [a_{k_1}, \ldots, a_{k_m}]$ denotes all Fourier Contour Embedding vectors in this frame, which each element in $\hat{\mathbf{a}_k}$ can be obtained by Fourier Transformation after discretizing continual contour $C(t)$ into N points sequence $C(\frac{n}{N})$.

$$a_k = \frac{1}{N} \sum_{n=1}^{N} C(\frac{n}{N}) e^{-2\pi i k \frac{n}{N}} \tag{3}$$

Each combination of a_k and $e^{2\pi i k}$ represents a circular motion with initial vector a_k and frequency k. Consequently, as shown in Fig. 1, we can regard the text contour as integration of circular motions with different frequency (pink circles in figure). Each pixel in text contour contains VTD-FCE vector $[u_{-k}, v_{-k} \ldots u_k, v_k, a]$, where u and v represent the real part and image part of Fourier Contour Embedding vector a_k, a donates scales to be activated. In our method, we set $k = 5$.

Our VTD-FCE method first resample the contour between ground truth points in a fixed number N to obtain dense point sequence. Then Fourier Transformation is adopted to get Fourier signature a_k with resampled contour points. Finally by integrating circular motions as shown in Fig. 1, we

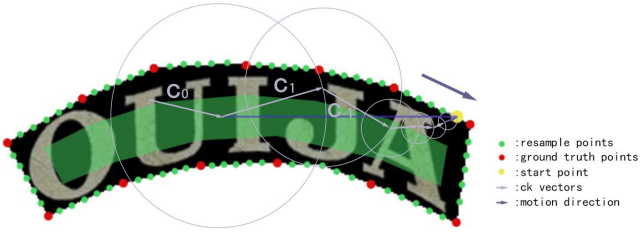

Fig. 1 Description of VTD-FCE. We first resample the origin ground truth contour points (red points in figure) to get dense points (green points in figure). Then Fourier Transformation is adopted to get Fourier signature a_k with resampled points. Last by integrating different circular motions (pink circles in figure), we can reconstruct the text contour.

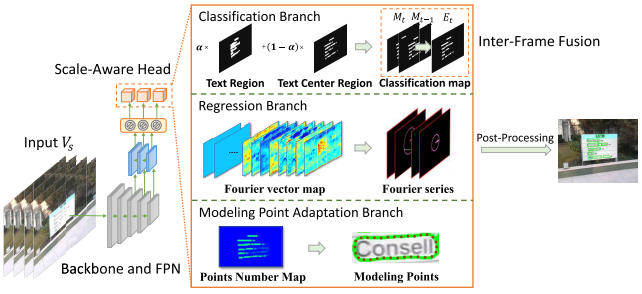

Fig. 2 The framework of VTD-FCENet. Video frames are sent into backbone and FPN to extract features which are fed through scale-aware head to detect texts. The scale-aware head activates different prediction head by automatically calculating size distribution of texts and analyzing scale category of dataset. Each head consists of three branches, which predicts text classification map with inter-frame fusion, Fourier series and modeling points number, respectively. After post-processing via inverse Fourier transform and non-maximum suppression, we obtain the final output.

can reconstruct the text contour.

Note that constrains on starting point, sampling direction and moving speed are utilized to make Fourier signature a_k unique. We set our starting point to be right most intersection point between the horizontal line through the center point and the text contour. Sampling direction is set in clockwise direction and moving speed is uniform.

A scale-aware mechanism is designed to adaptively select the scale of the feature output corresponding to the backbone network in the training stage based on the size of the data. During training, this module automatically calculates the size distribution of texts in the dataset and divides them into three categories based on the size of the text. We utilize different scales of feature output and different prediction heads for each of the three categories in the network, and adaptively select the scale based on the distribution of size ratios. When disribution proportion of the category is lower than a threshold θ, we freeze and remove the corresponding scale head to increase efficiency and reduce the impact of other scales. For the remaining scales, scale with the highest distribution proportion is supervised with the input samples of all sizes, while other scales are only supervised with their corresponding sizes.

2.2 VTD-FCENet for Real-Time Video Text Detection

Network Architectures. Equipped with VTD-FCE, we propose VTD-FCENet to achieve real-time video text detection. Different from FCENet [2] which only uses the same head for multi-scale outputs, we set separate scale-aware predictions head for each individual layer of feature output to better supervise the scale changes. Our VTD-FCENet consists of ResNet [14] as the backbone, FPN [15] as the neck, and three separate prediction heads. Different scale feature ouput of FPN will be fed into different prediction heads to predict text regions, text center regions, Fourier vectors and modling points number. The final detection results would be obtained through post-processing.

The prediction head consists of three branches, where the classification branch predicts the text region (TR) mask at the pixel level; the regression branch predicts the Fourier vectors of the contour of text instances; and the modeling point adaptation branch predicts the modeling point number used for post-processing based on the frame's complexity. Each branch contains three 3×3 convolutional layers and one 1×1 convolutional layer, and each of them is followed by a ReLU layer.

In addition, inter-frame fusion module is designed to exploit the correlation between adjacent frames in a video stream. We collect the predicted output mask M_{t-1} and M_t from adjacent frames with thresholds β_1 and β_2. At first, we filter the predicted mask from previous frame M_{t-1} by β_1 to obtain M'_{t-1}. Then, the filtered M'_{t-1} and predicted mask of the current frame M_t are combined and filtered by β_2 to get the enhanced prediction E_t in the current frame.

Ground-Truth Generation. In the classification branch, we use the method of [2] to obtain the text center region (TCR) of the mask to shrink the text by a factor of 0.3. In the regression branch, the Fourier vectors will be regressed in each pixel of the text contour. In the adaptive sample points task, we determine the sample points number based on the number of text instances present in the frame. We adopt a smaller sample points number when there are more text instances in the frame to maintain stable speed under different conditions.

Loss Function. The loss function of VTD-FCENet is $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{sam}$, where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$ and $\mathcal{L}_{sam}$ are the losses of the classification, regression and adaptive sample points branch, respectively.

For $\mathcal{L}_{cls}$, it consists of two components, i.e. $\mathcal{L}_{cls} = \mathcal{L}_{tr} + \mathcal{L}_{tcr}$, where $\mathcal{L}_{tr}$ and $\mathcal{L}_{tcr}$ are the cross-entropy losses of text region (TR) and text center region (TCR), respectively. To solve the sample imbalance problem, the OHEM method is used with the ratio 3:1 of negative samples to positive samples. For $\mathcal{L}_{reg}$, we minimize reconstructed text contours in the image space domain instead of predicted Fourier vectors. For $\mathcal{L}_{sam}$, we adopt the cross-entropy losses of predicted sample points number in text region to calculate.

Post Processing. The confidence of the predicted text contour C is obtained via weighted summation of the text region confidence C_{tr} and text center region confidence C_{tcr}, i.e. $C = \alpha C_{tr} + (1 - \alpha) C_{tcr}$. The typical value of α was set to 0.1 in our experiments. Then, the predicted output with high confidence would be utilized to reconstruct text con-

tours via inverse Fourier transform (IFT) and non-maximum suppression (NMS).

3. Experiment

Experimental evaluation of both detection accuracy (measured by precision P, recall R, and f-measure F) and inference speed (measured by frames per second fps) were conducted on three benchmark datasets for VTD tasks, including ICDAR 2013 Video, Minetto and YVT.

ICDAR 2013 Video [11] (frame size ranges from 720×480 to 1280×960) contains 13 training videos and 15 test videos, captured by 4 cameras in indoor and outdoor scenes. Minetto (frame size 640×480) [12] contains 5 videos of outdoor scenes. YVT [13] contains videos (frame size 1280×720) collected from youtube, where half is for training and the other is for testing.

3.1 Implementation Details

The backbone of model was initialized with the model pretrained on ImageNet. The optimizer uses stochastic gradient descent with the momentum of 0.9. The initialized learning rate is 0.001, which is reduced 0.8× every 100 epoches. Before training, we identify and remove such frames beforehand to avoid negative impact. In training stage, models for ICDAR 2013 and YVT are first pretrained on ICDAR 2015 and then finetuned on their own dataset. Since the Minetto dataset only have a testing set, we use the models trained on ICDAR 2013 for testing. In testing stage, thresholds of text region was set to 0.95 for ICDAR2013 and Minetto, 0.9 in YVT. Threshold of NMS in post-processing was set to 0.05.

3.2 Basic Evaluation

Both evaluations of detection accuracy and speed were conducted for VTD-FCENet on ICDAR 2013 Video, Minetto and YVT datasets, and the results indicate the effectiveness of VTD-FCE and VTD-FCENet for the real-time VTD task.

Evaluation of VTD-FCE. The VTD-FCE method is evaluated via comparison of a CNN-based detector without VTD-FCE and a detector with VTD-FCE, as shown in Fig. 3. It can be seen that the detected boundary produced by VTD-FCE fit text instances closely. It is worth mentioning that a prominent advantage of our VTD-FCE method is the ability to model irregular text. However, there are few irregular texts in existing public video text datasets, which cannot show our ability in this regard.

Our method still has limitation like lack of ability to solve the domain difference in samples. As shown in Table 1, the performance on YVT is not superior as other two datasets. That's because YVT consists of cartoons, albums which include a lot of synthtext and wordart while ICDAR 2013 Video and Minetto are both collected from natural scenes. Our model did not achieve adequate generalization ability to solve the domain shift problem. Besides, we didn't perform well on some slender and small texts. As shown in Fig. 4, our method can't detect text correctly, even can not detect anything in some situations. For the limitation and weakness of our method, we will develop them further in the future version.

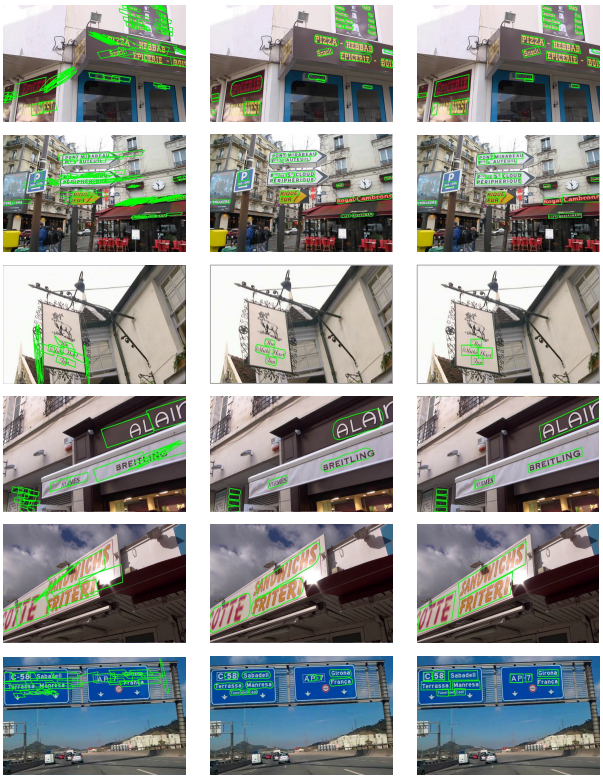

Fig. 3 Evaluation of VTD-FCE on ICDAR2013Video, Minetto and YVT, where the left column shows results of the detector without VTD-FCE, the middle column shows the results with VTD-FCE, and the right column shows the ground-truth.

Fig. 4 Limitation on VTD-FCE. VTD-FCE did not perform well on samples which include synthtext, wordart and some slender and small texts.

Ablation Study of VTD-FCENet. We conducted ablation studies of the proposed VTD-FCENet, shown in Table 2. We tested the performance among the scale-aware network, text region weighted sum, inter-frame fusion module and GPU inference acceleration, respectively. The results indicate the effectiveness of the components of VTD-FCENet to improve the accuracy and speed for the VTD task.

Speed Evaluation on HD videos. We also evaluated the speed of our method on videos with various resolutions. As shown in Table 3, our model can perform real-time detection on full HD resolution (1080p) videos, and even higher

Table 2 Albation Study of VTD-FCENet. "Net", "Wei", "Acc", "Fus" donates Scale-Aware optimized network, weighted sum, GPU acceleration, fusion module, respectively. "F" and "S" denote F-Measure (%) and Speed (fps).

Net	Wei	Acc	Fus	IC13		Mine		YVT	
				F	S	F	S	F	S
-	-	-	-	66.2	21.8	86.4	24.0	72.0	30.6
√	-	-	-	67.4	31.7	88.1	34.1	73.4	42.3
√	√	-	-	67.4	30.6	88.8	33.1	75.5	42.5
√	√	√	-	69.1	39.3	91.2	43.0	77.0	54.9
√	√	√	√	69.4	39.2	91.5	42.9	77.3	54.9

Table 3 Inference speed of VTD-FCENet on video with different resolution from ICDAR2013, Minetto, and YVT datasets.

Video Resolution	ICDAR2013	Minetto	YVT
SD (480p)	119fps	110fps	141fps
HD (720p)	75fps	60fps	88fps
Full HD (1080p)	40fps	33fps	44fps

frame rates of up to 60fps on HD resolution (720p) videos.

3.3 Comparison with Related Methods

We made extensive comparison with many related methods on ICDAR2013 Video, YVT and Minetto datasets, as shown in Table 1. For detection accuracy, the results illustrate that our VTD-FCENet obtains the best performance of F-measure on both ICDAR2013 and Minetto datasets, and obtain competitive performance on YVT dataset. For inference speed, our VTD-FCENet method not only obtains the highest speed, but also is the only one method that achieve real-time VTD on different datasets, even for HD videos.

We also made comparison with VTD-FCENet and our preliminary FCENet [2] that is originally designed for text detection in images. The result shows that directly using the FCENet method for VTD task is sub-optimal for detection accuracy due to the lack of inter-frame information. But we can see that benefiting from the FCE signature in Fourier domain, even the original FCENet obtains highest inference speed (over 30fps on YVT) among pervious methods, which show the potential of FCE for VTD. Therefore, based on FCE, we design scale-aware VTD-FCE and construct VTD-FCENet with a more lightweight architecture to obtain better detection accuracy and speed, and the results verify the effectiveness of our method.

4. Conclusion

This paper proposes a VTD-FCE method, which adaptively select the scale of text instances. Based on VTD-FCE, VTD-FCENet is constructed with inter-frame fusion. Experimental results on thee benchmark datasets show that our VTD-FCENet not only obtains SOTA or competitive detection accuracy, but also obtains real-time inference speed simultaneously.

Acknowledgements

Lingyu Liang was supported by the Fundamental Research Funds for the Central Universities, the Open Fund of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (MJUKF-IPIC202102) and the Science and Technology Program of Pazhou Lab. Tao Wang was supported by Fujian Provincial Natural Science Foundation General Project (2022J011112), Research Project of Fashu Foundation (MFK23001), The Open Program of The Key Laboratory of Cognitive Computing and Intelligent Information Processing of Fujian Education Institutions, Wuyi University (KLCCIIP2020202).

References

[1] W. Shao, R. Kawakami, and T. Naemura, "Anomaly detection using spatio-temporal context learned by video clip sorting," IEICE Tran. Inf. & Syst., vol.105, no.5, pp.1094–1102, 2022.
[2] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin, and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection," Proc. CVPR, pp.3123–3131, 2021.
[3] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo, "Bridging video-text retrieval with multiple choice questions," Proc. CVPR, pp.16167–16176, 2022.
[4] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas, and C. Jawahar, "Roadtext-1k: Text detection & recognition dataset for driving videos," Proc. ICRA, pp.11074–11080, 2020.
[5] P. Shivakumara, L. Wu, T. Lu, C.L. Tan, M. Blumenstein, and B.S. Anami, "Fractals based multi-oriented text detection system for recognition in mobile video images," Pattern Recognition, vol.68, pp.158–174, 2017.
[6] Y. Gao, X. Li, J. Zhang, Y. Zhou, D. Jin, J. Wang, S. Zhu, and X. Bai, "Video text tracking with a spatio-temporal complementary model," IEEE Trans. on Image Processing, vol.30, pp.9321–9331, 2021.
[7] H. Yu, Y. Huang, L. Pi, C. Zhang, X. Li, and L. Wang, "End-to-end video text detection with online tracking," Pattern Recognition, vol.113, 107791, 2021.
[8] W. Feng, F. Yin, X.-Y. Zhang, and C.-L. Liu, "Semantic-aware video text detection," Proc. CVPR, pp.1695–1705, 2021.
[9] L. Chen, J. Shi, and F. Su, "Robust video text detection through parametric shape regression, propagation and fusion," Proc. ICME, pp.1–6, 2021.
[10] L. Wang, J. Shi, Y. Wang, and F. Su, "Video text detection by attentive spatiotemporal fusion of deep convolutional features," Proc. ACM MM, pp.66–74, 2019.
[11] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L.G. Bigorda, S.R. Mestre, J. Mas, D.F. Mota, J.A. Almazàn, and L.P. De Las Heras, "ICDAR 2013 robust reading competition," Proc. ICDAR, pp.1484–1493, IEEE, 2013.
[12] R. Minetto, N. Thome, M. Cord, N.J. Leite, and J. Stolfi, "Snoopertrack: Text detection and tracking for outdoor videos," Proc. ICIP, pp.505–508, 2011.
[13] P.X. Nguyen, K. Wang, and S. Belongie, "Video text detection and recognition: Dataset and benchmark," Proc. WACV, pp.776–783, 2014.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. CVPR, 2016.
[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. CVPR, 2017.
[16] L. Wang, Y. Wang, S. Shan, and F. Su, "Scene text detection and tracking in video with background cues," Proc. ACM ICMR, pp.160–168, 2018.
[17] Y. Wang, L. Wang, and F. Su, "A robust approach for scene text detection and tracking in video," Proc. PCM, pp.303–314, 2018.