



Nice, France

October 21-25 2019

You Only Recognize Once: Towards Fast Video Text Spotting

Zhanzhan Cheng; **Jing Lu**; Yi Niu; Shiliang Pu
Hikvision Research Institute

Fei Wu
Zhejiang University

Shuigeng Zhou
Fudan University

lujing6@hikvision.com

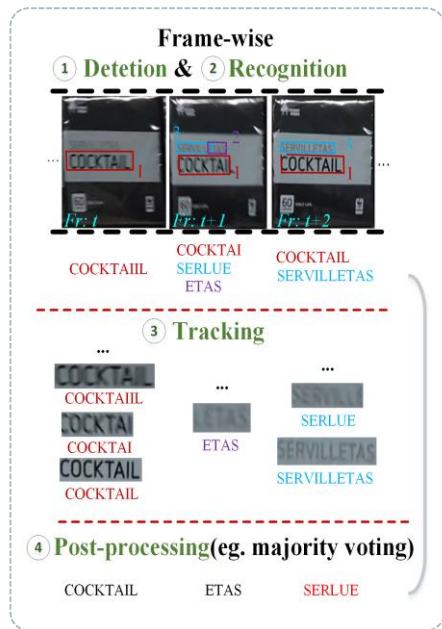
HIKVISION



YORO-Fast Video Text Spotting

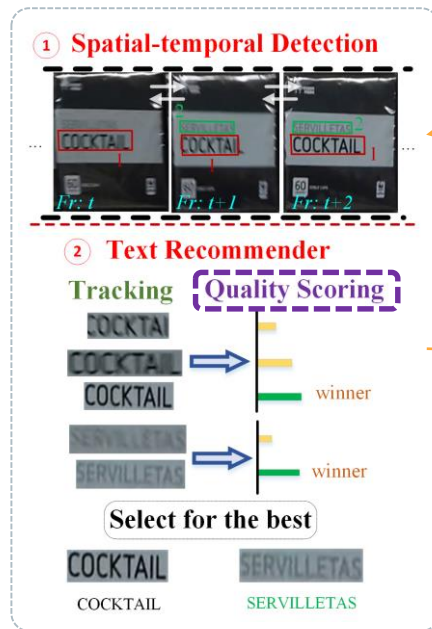
■ Motivation

● Problems of existing framework



- *unstable recognition results*
- *excessive computational cost*

● Advantages of our framework



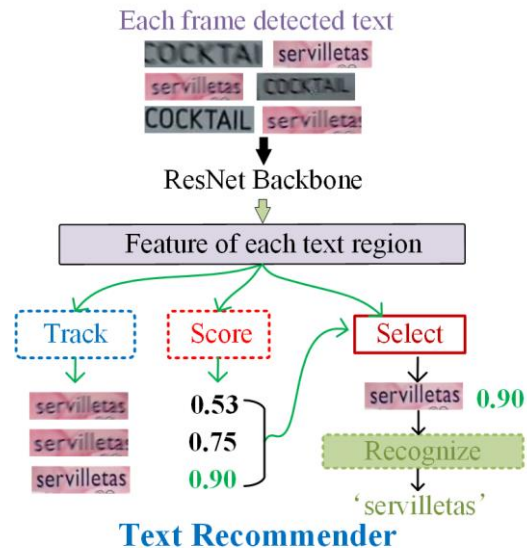
- *more robust results*
- *faster by recognizing once*

■ Contribution

● An unified two-stage framework YORO for fast video text spotting :

- *A self-attention based robust detector*
- *A novel text recommender for fast text recognition*

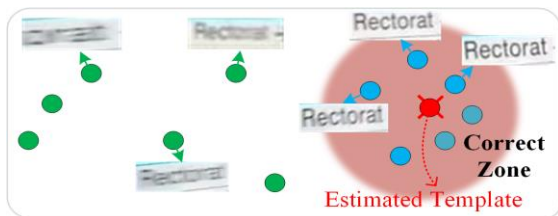
● Fast and robust recognition:



YORO-Fast Video Text Spotting

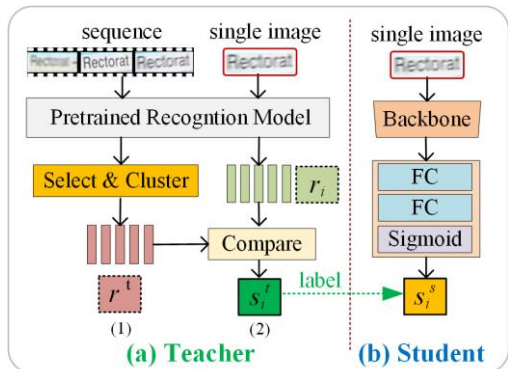
Key component

Mechanism of quality scoring network



$$r^t = kmeans(r_1^{cor}, r_1^{cor}, \dots, r_k^{cor})$$

$$s_i^t = \frac{r^t \odot r_i}{\|r^t\| * \|r_i\|}$$



(1) Teacher-Student Architecture

Experiments

Abalation

Methods	QSHR (IC13/IC15)	RCR (IC13/IC15)	FPS
PCW	74.55/75.83	66.06/66.32	4.52
HFP	75.32/76.34	68.30/68.56	x 80 faster 324.58
TR (\mathcal{L}_S)	77.89/79.69	68.89/69.41	
TR ($\mathcal{L}_S + \mathcal{L}_T$)	78.64/80.36	69.12/69.82	
TR ($\mathcal{L}_S + \mathcal{L}_R$)	81.23/83.03	69.92/70.69	
TR (\mathcal{L})	81.74/83.29	70.18/70.95	

(a) comparison with other frame selection methods

Method	✓	✓	✓	✓
D-BASE	✓	✓		
D-ST			✓	✓
TR (\mathcal{L}_S)	✓		✓	
TR		✓		✓
PRE_S	69.91	72.84	64.88	68.28
REC_S	54.34	61.73	61.54	67.21
$F-score$	61.15	66.83	63.17	67.74

(b) effectiveness of each module on IC15

Comparison with SoTA

Methods	REC	PRE	F-measure
Khare et al. [20]	41.40	47.60	44.30
Zhao et al. [58]	47.02	46.30	46.65
Shivakumara [42]	53.71	51.15	50.67
Yin et al. [55]	54.73	48.62	51.56
Wang et al. [52]	51.74	58.34	54.45
D-BASE	56.21	85.76	67.91
D-ST	60.23	81.45	69.25

(c) performance of detection on IC13

Method	$MOTP_R$	$MOTAR$	$ATAR$
Stradvision [18]	0.69	0.57	0.29
Deep2Text [18]	0.62	0.35	0.19
Wang et al. [53]	0.70	0.69	0.60
Ours	0.76	0.69	0.63

(d) performance of end-to-end on IC15

The Large Scale Video Text Dataset

Datasets	#scenarios	#videos	#frames	#instances	quality?
Merino [28]	4	-	-	-	
Minetto [30]	-	5	3599	8706	
IC13 [19]	7	28	15277	93934	✓
YVT [33]	-	30	13500	-	
IC15 [18]	7	49	27824	-	✓
LSVTD	22	100	66700	569300	✓

- ✓ Much larger scale(22 scenes)
- ✓ Multilingual text
- ✓ Release soon...