A Fast and Accurate Object Detector for Handwritten Digit String Recognition

Jun Guo, Wenjing Wei, Yifeng Ma and Cong Peng

School of Data Science & Engineering

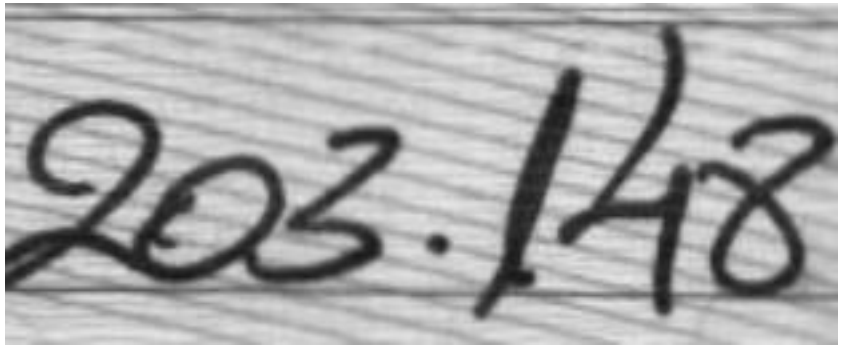
East China Normal University, China

jguo@cc.ecnu.edu.cn

Related work

- ✓ Traditional Model

  segment and recognize



HDSR

- ✓ Sequence-based Model

  Convolutional recurrent neural network (CRNN)
  CNN + RNN + connectionist temporal classification (CTC)

- ✓ Object Detection Model

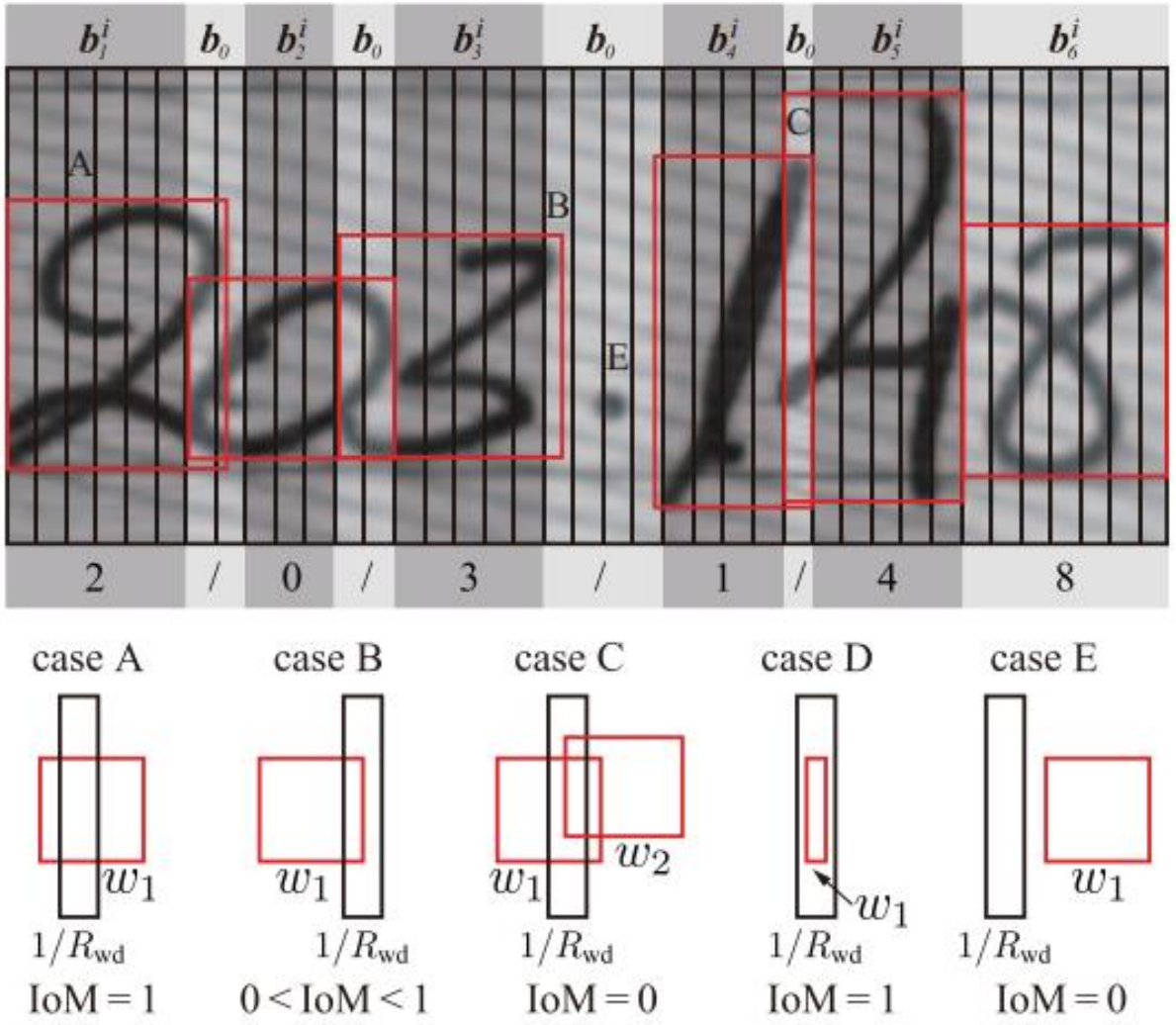
  Faster R-CNN
  YOLO

Existing problem

Anchor-based Model

ChipNet

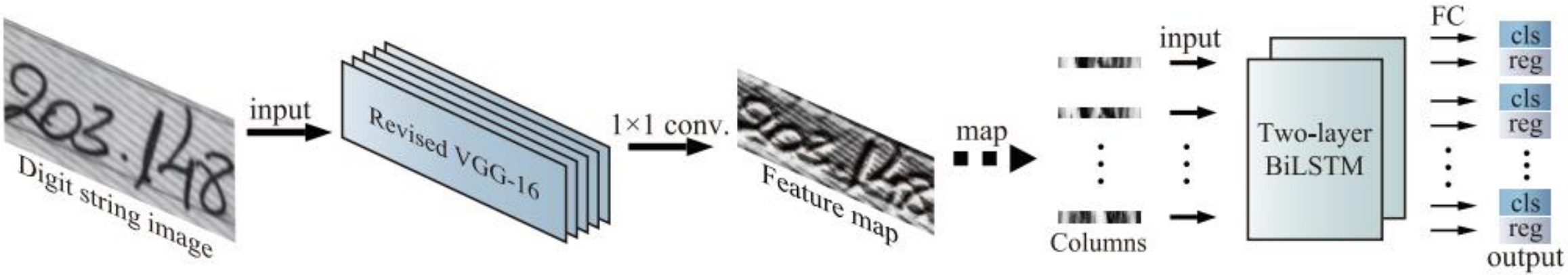
Encoding



Class	HDS5	CVL HDS	CAR-A	CAR-B
0	679,381	10,086	16,508	51,250
1	328,699	7,762	8,616	7,617
2	748,993	5,257	8,408	10,499
3	580,264	7,197	7,355	7,762
4	646,143	7,796	6,381	7,182
5	668,890	11,128	8,472	10,823
6	534,451	5,518	5,923	5,181
7	580,546	7,701	5,323	5,314
8	580,194	4,375	5,870	5,698
9	500,475	8,197	5,034	4,587
foreground	5,848,036	75,017	77,890	115,913
background	10,151,964	86,391	179,262	268,087
ratio	1 : 1.7	1 : 1.2	1 : 2.3	1 : 2.3

Class-balanced

The architecture of ChipNet



- ✓ **Anchor-free**
- ✓ **No RoI pooling**

Network design

Layer(type)	Configurations
1: FC	cls: $W^i/2 \times 11$, a: softmax reg: $W^i/2 \times 4$, a: sigmoid
2: BiLSTM	hidden units: $W^i/2$, size: 2048, ln
3: BiLSTM	hidden units: $W^i/2$, size: 2048
4: Reshape	–
5: **Convolution**	**f: 1**, k: 1×1, s: 1×1, p: same, a: relu, bn
6: Convolution$\times$3	f: 512, k: 3×3, s: 1×1, p: same, a: relu
7: Convolution$\times$3	f: 512, k: 3×3, s: 1×1, p: same, a: relu
8: Convolution$\times$3	f: 256, k: 3×3, s: 1×1, p: same, a: relu
9: **MaxPooling**	**k: 1×2, s: 1×2**
10: Convolution$\times$2	f: 128, k: 3×3, s: 1×1, p: same, a: relu
11: MaxPooling	k: 2×2, s: 2×2
12: Convolution$\times$2	f: 64, k: 3×3, s: 1×1, p: same, a: relu
13: Input	$W^i \times H$ grayscale images

Experiments on HDS5

Data	Length	Initial samples	Randomly select	
			Training set	Test set
R_{tablet}	2	13,405	10,000	800
	3	12,503	10,000	800
	4	11,289	10,000	800
	5	10,928	10,000	800
	total	48,125	40,000	3,200
S_{tablet}	2	87,968	25,000	500
	3	87,969	25,000	500
	4	87,969	25,000	500
	5	87,969	25,000	500
	total	351,875	100,000	2,000
S_{MNIST}	2	50,000	15,000	700
	3	50,000	15,000	700
	4	50,000	15,000	700
	5	50,000	15,000	700
	total	200,000	60,000	2,800
HDS5	2	151,373	50,000	2,000
	3	150,472	50,000	2,000
	4	149,258	50,000	2,000
	5	148,897	50,000	2,000
	total	600,000	**200,000**	**8,000**

Details of HDS5

Model	IoU	Accuracy	mAP	FPS
Faster R-CNN*	0.5	94.67	99.18	5
	0.7	94.36	99.11	
	0.9	91.92	97.72	
YOLOv3-tiny*	0.5	95.57	99.26	**257**
	0.7	95.25	99.09	
	0.9	92.37	97.80	
ChipNet	0.5	**99.78**	**99.94**	219
	0.7	99.65	99.89	
	0.9	98.59	99.62	

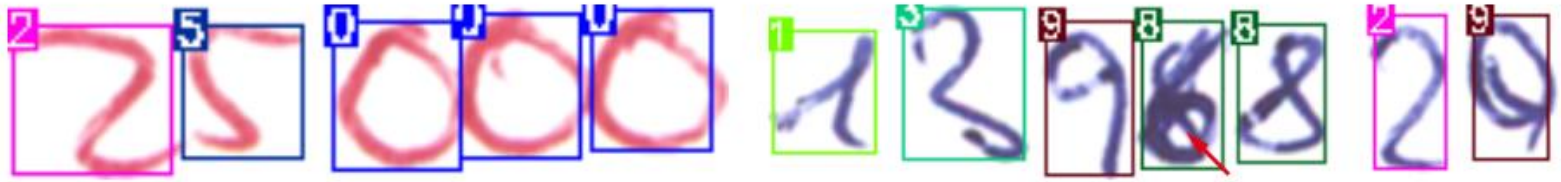
The results of the three detectors on HDS5

Experiments on benchmarks

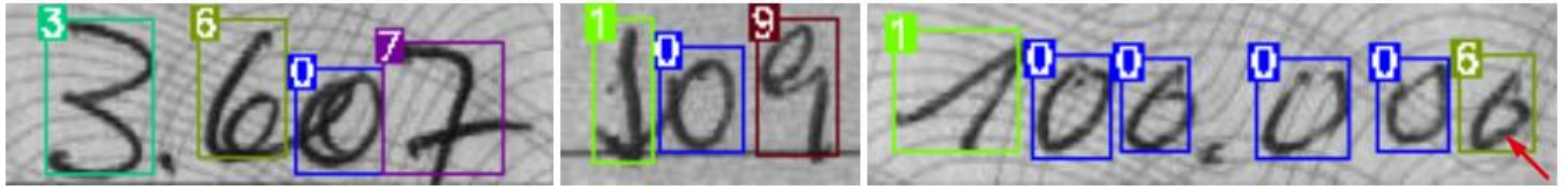
Method	CVL HDS	CAR-A	CAR-B
Tébessa I [24]	59.30	37.05	26.62
Tébessa II [24]	61.23	39.72	27.72
Shanghai [24]	48.93	49.50	28.09
Singapore [24]	50.40	52.30	59.30
Pernambuco [24]	58.60	78.30	75.43
Beijing [24]	85.29	80.73	70.13
Saabni [8]	-	85.80	
CRNN [9]	26.01	88.01	89.79
RNN-CTC [10]	27.07	89.75	91.14
Faster R-CNN*	69.97	74.97	76.22
YOLOv3-tiny*	65.86	72.51	76.13
ChipNet	**91.91**	**92.36**	**92.89**

Visualization results

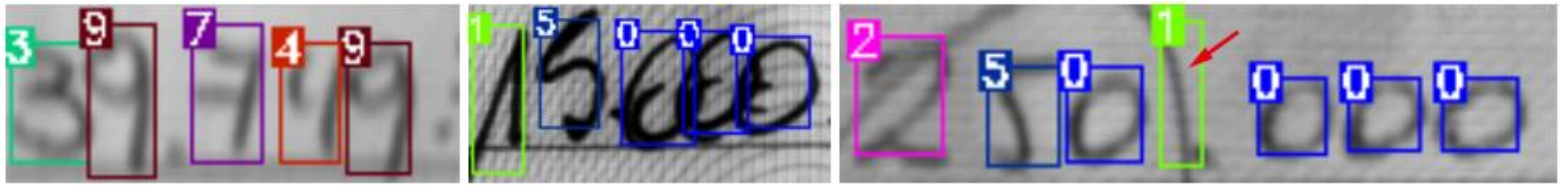
(a) Strings ‘052’, ‘1189’, ‘36995’ and ‘2154’ from HDS5.



(b) Strings ‘25000’ and ‘1396829’ from CVL HDS.



(c) Strings ‘3607’, ‘109’ and ‘100000’ from CAR-A.



(d) Strings ‘39749’, ‘15000’ and ‘250000’ from CAR-B.

Ablation study

Model	HDS5	CVL HDS	CAR-A	CAR-B
Model 1	98.12 298	89.86 272	90.06 328	90.85 325
Model 2	99.16 245	90.92 226	91.21 289	91.97 291
Model 3	99.67 185	91.52 172	91.65 217	92.03 221
Model 4	97.56 216	88.79 205	90.12 259	90.56 262
ChipNet	99.78 219	91.91 201	92.36 267	92.89 262

Conclusions

- ✓ A new object detector for HDSR
- ✓ An effective encoding method
- ✓ No region proposals, anchors and RoI pooling
- ✓ An accuracy of 99.78% on HDS5
- ✓ A real-time speed