NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

GoMatching: A Simple Baseline for Video Text Spotting via Long and Short Term Matching

Haibin He[1], Maoyuan Ye[1], Jing Zhang[1], Juhua Liu[1], Bo Du[1], Dacheng Tao[2]

[1] Wuhan University, China [2] Nanyang Technological University, Singapore

{haibinhe, yemaoyuan, liujuhua, dubo}@whu.edu.com

jingzhang.cv, dacheng.tao}@gmail.com

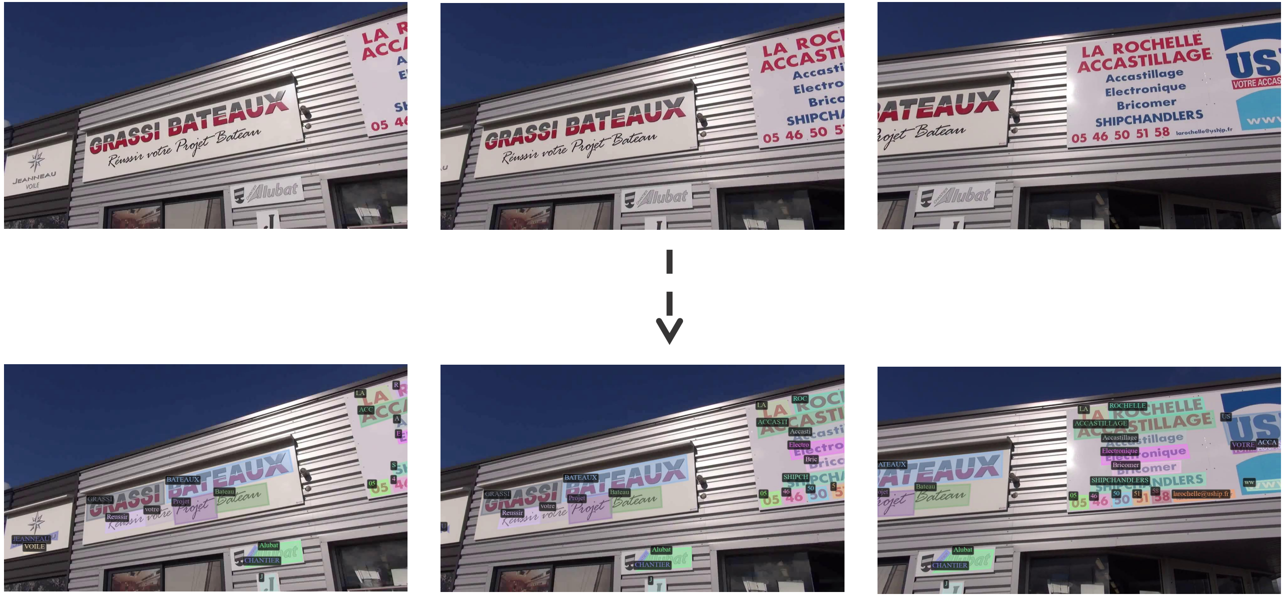
BackGround

Text spotting

Image

detection + recognition

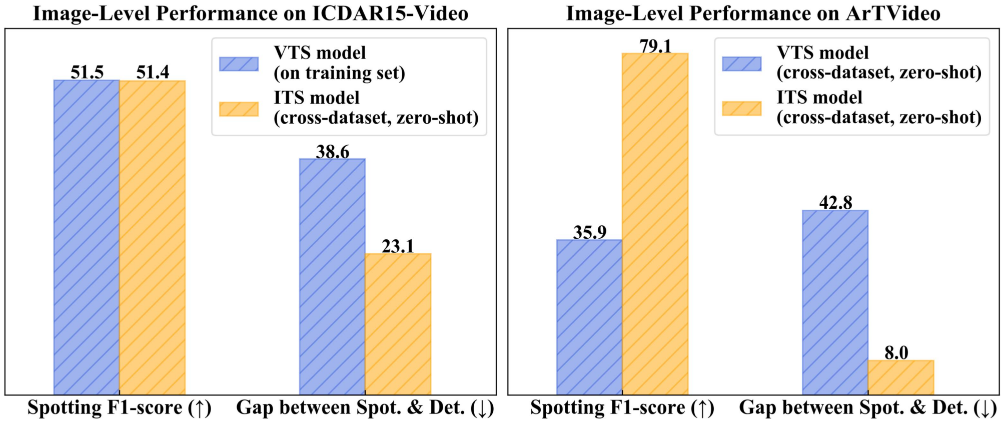
Video



detection + recognition + tracking

Motivation

- Current state-of-the-art video text spotter has a main bottleneck: **the limited recognition capability.**
- Directly adopting a frozen image text spotter **leads to low confidence** and consequently **a relatively low Recall** on video data. Moreover, the image text spotter **lacks the capability to track** the text instances across frames.
- Since the **scarcity of curved text instances** within existing video text spotting datasets, evaluating the performance of recognizing curved text is still infeasible.



How to effortlessly turn an image text spotter into an expert on video ?

Figure. 'Gap between Spot. & Det.': the gap between spotting and detection F1-score. The **larger** the gap, the **poorer** the recognition ability. Compared to the Image Text Spotting (ITS) model, the Video Text Spotting (VTS) model presents unsatisfactory text spotting F1-scores, which lag far behind its detection performance, especially on ArTVideo with curved text.

Contributions

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

- We **identify the limitations in current VTS methods** and **propose a novel and simple baseline**, which leverages an off-the-shelf image text spotter with a strong customized tracker.

- We introduce **the rescoring mechanism** and **long-short term matching module** to adapt image text spotter to video datasets.

- We establish the **ArTVideo test set** for addressing the absence of curved texts in current video datasets and evaluating the text spotters on **videos with arbitrary-shape text**.

Methodology

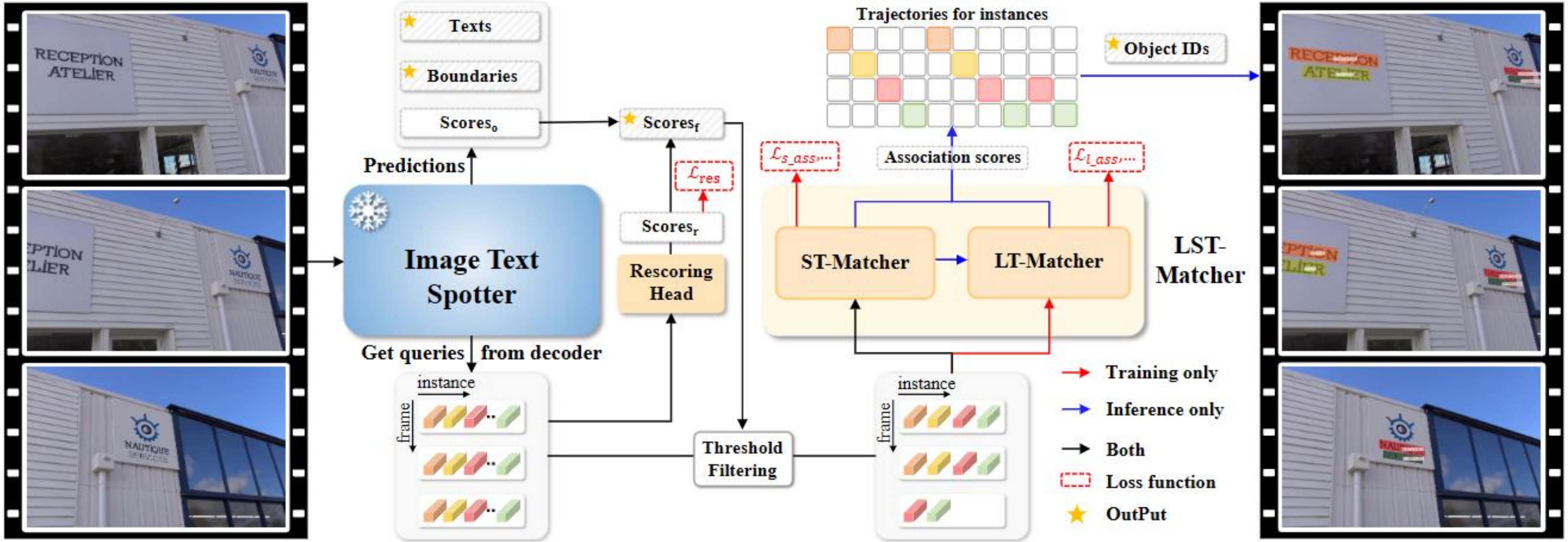


Figure 2: **The overall architecture of GoMatching.** The frozen image text spotter provides text spotting results for frames. The rescoring mechanism considers both instance scores from the image text spotter and a trainable rescoring head to reduce performance degradation due to the domain gap. Long-short term matching module (LST-Matcher) assigns IDs to text instances based on the queries in long-short term frames. The yellow star sign '★' indicates the final output of GoMatching.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Rescoring Mechanism

confidences output by frozen ITS model:

$$C_o^t = \{c_{o1}^t,\ c_{o2}^t, \ldots,\ c_{op}^t\}$$

t means the t-th frame,

confidences output by Rescoring head:

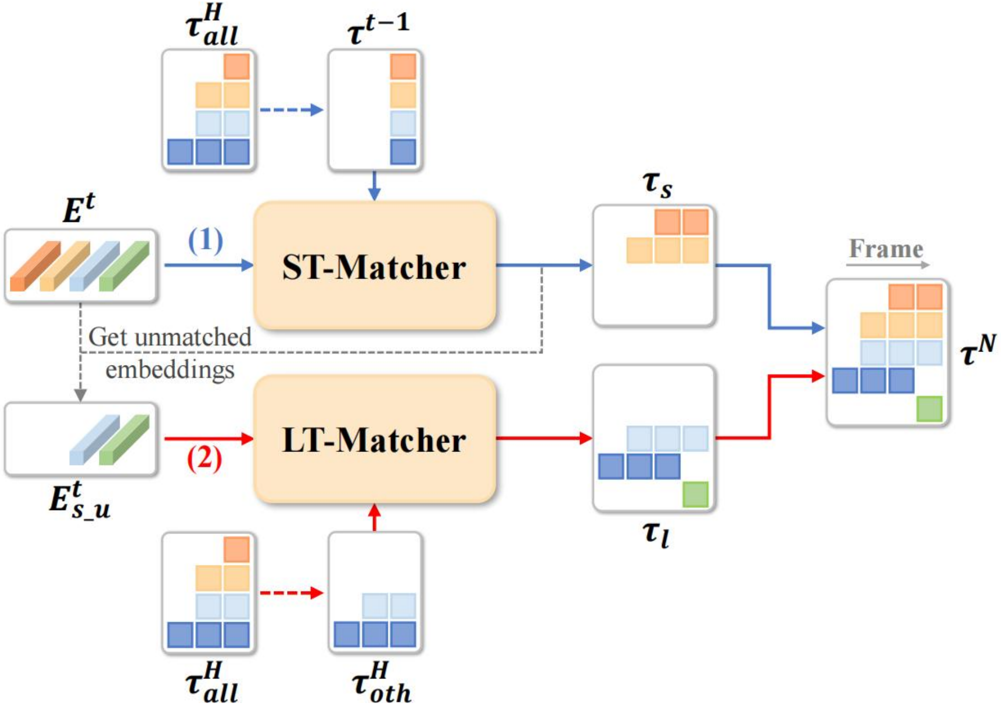
p means the num of queries

$$C_r^t = \{c_{r1}^t,\ c_{r2}^t, \ldots,\ c_{rp}^t\}$$

final confidences decided by score fusion operation:

$$C_f^t = \{c_{f1}^t = max(c_{o1}^t, c_{r1}^t),\ c_{f2}^t = max(c_{o2}^t, c_{r2}^t), \ldots,\ c_{fp}^t = max(c_{op}^t, c_{rp}^t)\}$$

LST-Matcher



① ST-Matcher first associates the detected instances with trajectories in previous frames as denoted by blue lines.

② LT-Matcher then associates the remaining unmatched instances by utilizing other trajectories in history frames as denoted by red lines.

Methodology

Training Loss

ReScoring Loss:
$$\mathcal{L}_{res} = \sum_{1}^{N} [-\mathbb{1}_{\{c_i \neq \varnothing\}} \alpha (1 - \hat{p}_{\hat{\sigma}(i)}(c_i))^{\gamma} \log(\hat{p}_{\hat{\sigma}(i)}(c_i)) - \mathbb{1}_{\{c_i = \varnothing\}} (1-\alpha)(\hat{p}_{\hat{\sigma}(i)}(c_i))^{\gamma} \log(1 - \hat{p}_{\hat{\sigma}(i)}(c_i))]$$

Long-Short Association Loss:
$$\mathcal{L}_{asso} = \mathcal{L}_{s\_bg} + \mathcal{L}_{l\_bg} + \sum_{\hat{\tau}_k} (\mathcal{L}_{s\_ass} + \mathcal{L}_{l\_ass})$$

$$\mathcal{L}_{s\_ass}(E^S, \hat{\tau}_k) = -\sum_{t=2}^{T} \log P_{s_a}(\hat{\alpha}_k^t | e^t_{\hat{\alpha}_k^t}, E^{S_t})$$

$$\mathcal{L}_{l\_ass}(E^L, \hat{\tau}_k) = -\sum_{w} \sum_{t=1}^{T} \log P_{l_a}(\hat{\alpha}_k^t | E^w_{\hat{\alpha}_k^w}, E^L)$$

$$\mathcal{L}_{s\_bg}(E^S) = -\sum_{j:\nexists \hat{\alpha}_k^t = j} \sum_{t=2}^{T} \log P_{s_a}(\alpha^t = \emptyset | e_j^t, E^{S_t})$$

$$\mathcal{L}_{l\_bg}(E^L) = -\sum_{w=1}^{T} \sum_{j:\nexists \hat{\alpha}_k^w = j} \sum_{t=1}^{T} \log P_{l_a}(\alpha^t = \emptyset | E_j^w, E^L)$$

Overall Loss:
$$\mathcal{L} = \lambda_{res}\mathcal{L}_{res} + \lambda_{asso}\mathcal{L}_{asso}$$

Experiments

Table 1: **Comparison results with SOTA methods on four distinct datasets.** '†' denotes that the results are collected from the official competition website. '*': we use the officially released model for evaluation. 'M-ME' indicates whether multi-model ensembling is used. 'Y' and 'N' stand for yes and no. The best and second-best results are marked in **bold** and underlined, respectively.

(a) **Results on ICDAR15-video.**

Method	MOTA (↑)	MOTP (↑)	IDF1 (↑)
HIK_OCR [9]	52.98	74.88	61.85
CoText [11]	58.94	74.53	71.66
TransDETR [12]	60.96	74.61	72.80
h&h_lab†	63.76	77.78	71.08
GOCR Offline†	63.05	74.31	76.95
CoText(Kuaishou_MMU)†	66.96	76.55	74.24
GoMatching (size:800) (ours)	68.51	77.52	76.59
GoMatching (size:1000) (ours)	**72.04**	**78.53**	**80.11**
GoMatching (size:1440) (ours)	70.52	78.25	78.70

(b) **Results on BOVText.**

Method	MOTA (↑)	MOTP (↑)	IDF1 (↑)
EAST + CRNN [10]	-79.3	76.3	6.8
PSENet + CRNN [10]	-17.0	79.2	31.3
DB + CRNN [10]	-13.2	81.3	38.8
TransVTSpotter [10]	-1.4	82.0	43.6
CoText [11]	11.4	80.3	48.3
GoMatching (ours)	**52.9**	**87.2**	**62.6**

(c) **Results on DSText.**

Method	M-ME	MOTA (↑)	MOTP (↑)	IDF1 (↑)
TransDETR+HRNet†	Y	-28.58	80.36	26.20
SCUT-MMOCR-KS†	Y	-27.47	76.59	43.61
TextTrack†	Y	-25.09	74.95	26.38
abcmot†	Y	5.54	74.61	24.25
DA†	Y	10.51	78.97	53.45
TencentOCR†	Y	22.44	**80.82**	**56.45**
TransDETR [12]*	N	-22.63	79.73	26.43
GoMatching (ours)	N	**22.83**	80.43	46.09

(d) **Results on ArTVideo.**

Method	MOTA (↑)	MOTP (↑)	IDF1 (↑)
ArTVideo Tracking			
TransDETR [12]	54.2	67.9	70.4
GoMatching (ours)	**67.2**	**81.3**	**75.8**
ArTVideo End-to-End Spotting			
TransDETR [12]	2.8	69.7	49.3
GoMatching (ours)	**68.8**	**82.9**	**78.5**
ArTVideo-Curved Tracking			
TransDETR [12]	4.4	60.5	50.2
GoMatching (ours)	**59.5**	**76.3**	**73.5**
ArTVideo-Curved End-to-End Spotting			
TransDETR [12]	-66.7	-61.9	26.9
GoMatching (ours)	**56.8**	**78.0**	**73.9**

Experiments

Experiments

Table 2: Impact of difference components in the proposed GoMatching. 'Query' indicates that LST-Matcher employs the queries of high-score text instances for association, otherwise RoI features. Column 'Scoring' indicates the employed scoring mechanism, in which 'O' means using the original scores from DeepSolo, 'R' means using the scores recomputed by the rescoring head, and 'F' means using the fusion scores obtained from the rescoring mechanism.

Index	Query	Scoring	LT-Matcher	ST-Matcher	MOTA (↑)	MOTP (↑)	IDF1 (↑)
1		O	✓		66.20	78.52	75.07
2	✓	O	✓		67.22	78.54	76.12
3	✓	R	✓		68.47	78.29	77.09
4	✓	F	✓		68.80	78.24	77.41
5	✓	F		✓	69.40	**78.34**	73.60
6	✓	F	✓	✓	**70.52**	78.25	**78.70**

Experiments

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Table 4: Ablation studies on the number of frames (T) for long-term association in LT-Matcher, and the max number of history frames in tracking memory bank is $H = T - 1$). Experiments are conducted on ICDAR15-video and the best results are marked in **bold**.

Number T	MOTA (↑)	MOTP (↑)	IDF1 (↑)
$T = 32$	70.13	78.07	78.24
$T = 16$	70.33	78.25	78.60
$T = 8$	70.44	78.25	**78.70**
$T = 6$	**70.52**	78.25	**78.70**
$T = 4$	70.51	**78.27**	78.16

Table 5: **Results of different score fusion strategies on ICDAR5-video.** 'Mean', 'Geo-mean', and 'Maximum' denote the arithmetic mean, geometric mean, and the maximum score fusion strategies, respectively. The best results are highlighted in **bold**.

Strategy	MOTA (↑)	MOTP (↑)	IDF1 (↑)
Mean	70.46	78.38	78.29
Geo-mean	70.29	**78.39**	78.26
Maximum	**70.52**	78.25	**78.70**

Table 6: **Results of using different image sizes on ICDAR15-video.** 'Size' means the size of the shorter side of the input image during inference. The best results are highlighted in **bold**.

Method	MOTA (↑)	MOTP (↑)	IDF1 (↑)	FPS (↑)
TransDETR(Size: 800)	60.96	74.61	72.80	12.69
GoMatching(Size: 800)	68.51	77.52	76.59	**14.41**
GoMatching(Size: 1000)	**72.04**	**78.53**	**80.11**	10.60

Table 7: **Comparison between TransDETR and GoMatching.** 'T-Para.' and 'A-Para.' denote the number of all parameters and the trainable parameters in each model, respectively.

Method	#T-Para. (M)	#A-Para. (M)
TransDETR	39.35	39.58
GoMatching	32.79	75.38

Thank you !

Paper

Code