ICDAR 2021 Competition on Scene Video Text Spotting

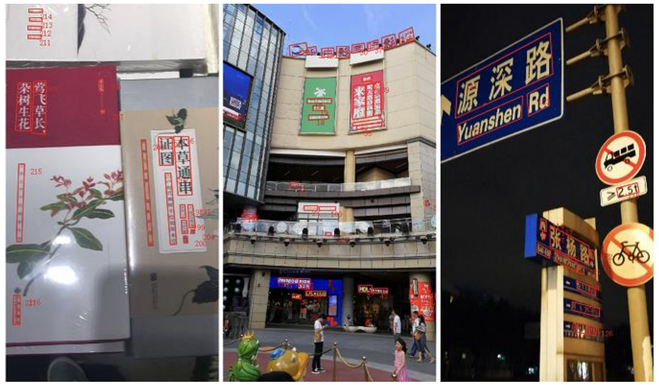
Zhanzhan Cheng; Jing Lu; Baorui Zou; Shuigeng Zhou; Fei Wu

Zhejiang University, Hikvision Research Institute, Fudan University

DAVAR Lab — Document Audio Video Analysis & Recognition LAB

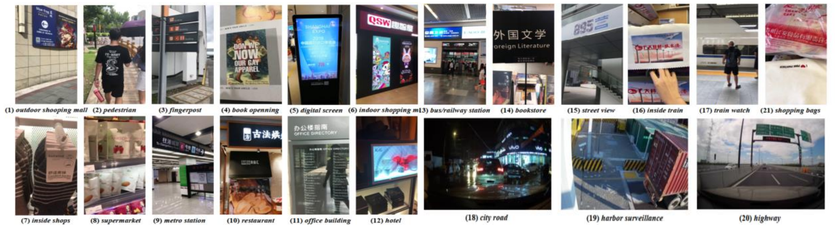
Introduction

- **Problems of SVTS Research**
  - Face various environmental interferences (e.g., camera shaking, motion blur and immediate illumination changing etc.) and meet the real-time response requirement
  - Existing datasets are too small to promote the area study
  - The lack of uniform evaluation metrics and benchmarks



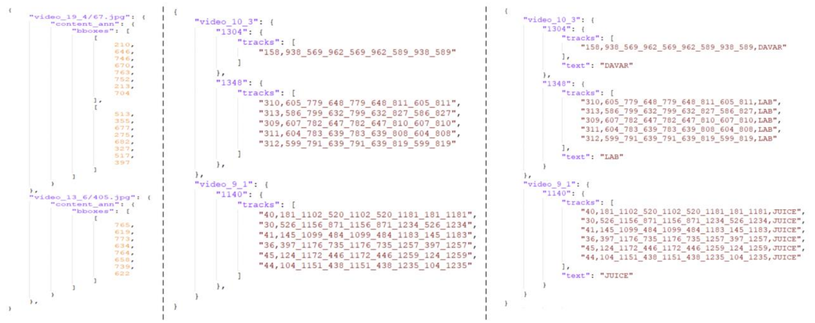
- **Competition Characteristics**
  - The video text dataset is further extended from LSVTD [1], containing 129 video clips from 21 real-life scenarios.
    - ✓ More accurate annotations
    - ✓ A general dataset with large range of scenarios
    - ✓ Video clips are overwhelming of low-quality images caused by blurring, perspective distortion, motion inferences etc.
  - Three specific tasks are proposed: video text detection, tracking and the end-to-end recognition
  - Provides the motivation, dataset description, task definition, evaluation metrics, results of submitted methods and their discussion.

Dataset



- **Characteristics**
  - Large scale and diversified scenes: 21 different scenes including 13 indoor and 8 outdoor scenes
  - Collected with different kinds of video cameras: mobile phone, HD camera, Car-DVR camera
  - Different difficulty levels: Hard, Medium and Easy
  - Multilingual instances: alphanumeric and non-alphanumeric
- **Dataset Split**: 71, 18 and 40 videos for training, validation and testing set, respectively.
- **Annotations**: *Polygon coordinate – unique identification – Language – Quality – Transcripts*

Tasks



Submited json files for Task 1, 2 & 3.

- **Task 1 – Video Text Detection**
  - $Recall_d$, $Precison_d$ and $F\text{-}score_d$ as the evaluation metrics
- **Task 2 – Video Text Tracking**
  - ATA_t, $MOTA_t$ and $MOTP_t$ are used as the evaluation metrics [3]
- **Task 3 – End2End Video Text Spotting**
  - Sequence-level evaluation protocols are proposed to evaluate the end-to-end performance, i.e., $Recall_s$, $Precision_s$, $F\text{-}score_s$ are used as evaluation metrics [2]

46 valid submissions from 24 teams from both research communities and industries for the three tasks

Submissions

User ID	Rank	F-score$_d$	Precison$_d$	Recall$_d$	Affiliations
tianqihenhao	1	0.8502	0.8561	0.8444	TEG, Tencent
wfeng	2	0.8159	0.8787	0.7615	IA, CAS
DXM-DI-AI -CV-TEAM	3	0.7665	0.8253	0.7155	DuXiaoman Financial
tangyejun	4	0.7582	0.8088	0.7136	*
wangsibo	5	0.7522	0.8377	0.6825	*
weijiawu	6	0.7298	0.7508	0.7098	Zhejiang University
yeah0110	7	0.7276	0.7314	0.7238	*
BOE_AIoT_CTO	8	0.7181	0.7133	0.7229	BOE
colorr	9	0.7172	0.7101	0.7245	*
qqqyd	10	0.7140	0.7045	0.7238	*
yucheng3	11	0.6749	0.8622	0.5544	University of Chinese Academy of Sciences
superboy	12	0.6704	0.8336	0.5607	*
seunghyun	13	0.6219	0.6897	0.5663	NAVER corp
hanquan	14	0.5881	0.6252	0.5552	*

Tab 1. Results of Video Text Detection

User ID	Rank	ATA$_t$	MOTA$_t$	MOTP$_t$	Affiliations
tianqihenhao	1	0.5372	0.7642	0.8286	TEG, Tencent
DXM-DI-AI -CV-TEAM	2	0.4810	0.6021	0.8017	DuXiaoman Financial
panda12	3	0.4636	0.7009	0.8277	IA, CAS
lzneu	4	0.3812	0.5647	0.8198	*
wangsibo	5	0.3778	0.5657	0.8200	*
yucheng3	6	0.3116	0.5605	0.8203	University of Chinese Academy of Sciences
tangyejun	7	0.2998	0.5027	0.8196	*
yeah0110	8	0.2915	0.4811	0.8218	*
sabrina_lx	9	0.2436	0.3757	0.7667	*
seunghyun	10	0.1415	0.2183	0.6949	NAVER corp
enderloong	11	0.0918	0.1820	0.7520	*
tiendv	12	0.0676	0.2155	0.7439	University of Information Technology
weijiawu	13	0.0186	0.1530	0.7454	Zhejiang University

Tab 2. Results of Video Text Tracking

User ID	Rank	F-score$_s$	Precision$_s$	Recall$_s$	ATA$_s$	MOTA$_s$	MOTP$_s$	Affiliations
tianqihenhao	1	0.5308	0.6655	0.4414	0.4549	0.5913	0.8421	TEG, Tencent
DXM-DI-AI -CV-TEAM	2	0.4755	0.6435	0.3770	0.4188	0.4960	0.8142	DuXiaoman Financial
panda12	3	0.4183	0.5243	0.3479	0.3579	0.5179	0.8427	IA, CAS
lzneu09	4	0.3007	0.3611	0.2576	0.2737	0.4255	0.8330	Northeastern University
yucheng3	5	0.2964	0.3506	0.2567	0.2711	0.4246	0.8332	University of Chinese Academy of Sciences
tangyejun	6	0.2284	0.2527	0.2084	0.2121	0.3676	0.8337	*
tiendv	7	0.0813	0.1402	0.0572	0.0802	0.0887	0.7976	University of Information Technology
enderloong	8	0.0307	0.0239	0.0429	0.0357	0.0159	0.7813	*
colorr	9	0.0158	0.0085	0.1225	0.0146	0.0765	0.8498	*
weijiawu3	10	0.0077	0.0041	0.0550	0.0088	-0.1530	0.7670	Zhejiang University
BOE_AIoT_CTO	11	0.0000	0.0000	0.0000	0.0000	-0.0003	0.0000	BOE

Tab 3. Results of End2End Video Text Spotting

Discussion

- **Video text detection task**
  - Most participants employ the semantic-based Mask R-CNN framework to capture regular and irregular text instance
  - TencentOCR team achieves the best score
- **Video text tracking task**
  - Most methods focus on the trajectory estimation
  - Tencent-OCR team achieves the best score
- **End2End Video Text Spotting**
  - A pre-trained general model is important in many method

The overall performance is low, and large improving space is existing for this research topic

Organization

- **Schedule of the SVTS competition**:
  - ✓ 5 January 2021: Registration begin. Training & validation datasets are available for downloads.
  - ✓ 1 March 2021: Submissions begin. Test data is released.
  - ✓ 31 March 2021: Registration deadline.
  - ✓ 11 April 2021: Submissions deadline of all the tasks.
- **Maintained on Codalab web**: *https://competitions.codalab.org/competitions/27667*

Scan Me for Paper (ICDAR-2021)

References

① *Cheng, Z., Lu, J., Niu, Y., Pu, S., Wu, F., Zhou, S.: You Only Recognize Once: Towards Fast Video Text Spotting. In: ACM MM. pp. 855{863 (2019)*

② *Cheng, Z., Lu, J., Zou, B., Qiao, L., Xu, Y., Pu, S., Niu, Y., Wu, F., Zhou, S.: FREE: A Fast and Robust End-to-End Video Text Spotter. IEEE Transactions on Image Processing 30, 822{837 (2020)*

③ *Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. In: ICDAR. pp. 1484-1493 (2013)*