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

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

- Task 1 – Video Text Detection
  - Recall, Precision, and F-score, as the evaluation metrics
- Task 2 – Video Text Tracking
  - ARA, MOTA, and MOTP, 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, Precision, F-score, are used as evaluation metrics [2]

Submissions

- 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
  - TencentOCR team achieves the best score
- End2End Video Text Spotting
  - A pre-trained general model is important in many methods

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

References