Supplarmentary Material for "MMVText"

Anonymous Author(s)
Affiliation
email

A Appendix

A.1 MMVText Metrics

The proposed MMVText mainly includes two tasks: (1) Video Text Tracking. (2) End to End Text Spotting in Videos. $MOTP$ (Multiple Object Tracking Precision) [1], $MOTA$ (Multiple Object Tracking Accuracy) and IDF_1 [2, 7] as the three important metrics are used to evaluate task1 (text tracking) for MMVText. Following the previous works [3, 6], MMVText evaluates text tracking methods in video and compare their performance with the MOTA and MOTP, which are given by:

$$MOTP = \frac{\sum_{i,t}(1 - d_t^i)}{\sum_t c_t}\,, \tag{1}$$

where c_t denotes the number of matches found for time t. For each of these matches, calculate the iou d_t^i between the object o^i and its corresponding hypothesis. It shows the ability of the tracker to estimate precise object positions. MOTA is calculated as follows:

$$MOTA = 1 - \frac{\sum_t(m_t + fp_t + mme_t)}{\sum_t g_t}\,, \tag{2}$$

where m_t, fp_t and mme_t are the number of misses, false positives, and mismatches, respectively. g_t is the number of objects present at time t. It shows the tracker's performance at detecting objects and keeping their trajectories, independent of the precision of the location. Besides, ID_{F1} as the new metrics for tracking is calculated as follows:

$$ID_{tp} = \sum_h \sum_t m(h, o, \triangle_t, \triangle_s)\,, \tag{3}$$

$$ID_{F1} = \frac{2ID_{tp}}{2ID_{tp} + ID_{fp} + ID_{fn}}\,, \tag{4}$$

where $\triangle_t$ and $\triangle_s$ refer to time matching and space location matching, respectively. ID_{tp}, ID_{fp} and ID_{fn} refer to true positive, false positive and false negative of matching ID.

In Task2 (End to End Text Spotting in Videos), we modify the $MOTA$, $MOTA$ and ID_{F1} to $MOTA_T$, $MOTA_T$ and TID_{F1}, the only difference is the matches need to meet the correct of the recognition result, since the recognition result is more important than tracking and localization. More specifically, TID_{F1} is calculated as follows:

$$TID_{tp} = \sum_h \sum_t m(h, o, \triangle_t, \triangle_s, \triangle_r)\,, \tag{5}$$

$$TID_{F1} = \frac{2TID_{tp}}{2TID_{tp} + TID_{fp} + TID_{fn}}\,, \tag{6}$$

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

Figure 1: **The Real Application Tasks Link to MMVText**. (a) Video Understanding, automatically describing visual content with natural language. (b) Video Caption Translation, extremely helpful for people who travel abroad and video-sharing websites such as YouTube. (c) Video Retrieval, accurate semantic information for text in videos can promote video retrieval.

where $\triangle_t$, $\triangle_s$ and $\triangle_r$ refer to time matching, space location matching and recognition result matching. And h and o denote hypothesis and true text trajectory with recognition result. The match of h and o is a true positives of text ID (*i.e.,* TID_{tp}) when these conditions (*i.e.,* $\triangle_t$, $\triangle_s$ and $\triangle_r$) are met. Similarly, false positive (*i.e.,* TID_{fp}) and false negative (*i.e.,* TID_{fn}) of text ID can be obtained for TID_{F1} calculation.

A.2 Link to Real Applications

In this section, we show that the practicability of the proposed MMVText, not a toy benchmark, which can promote other video-and-text application research.

Video Understanding. As shown in Figure. 1 (a), the example concerning the task of describing video with natural language is from MSR-VTT [10], and there has been increasing interest in video understanding [9, 4]. However, video description with only visual information is difficult and limited, even for a human. For the annotation of the sample video, *i.e., "A man in a blue suit and purple tie discusses millennial investing fear"*, we can not learn the information of "millennial investing fear" from the visual information in the video. By comparison, caption and scene texts in the video contain accurate information of *"millennial investing fear"*, which can help the model to describe the video better. We argue that the same as general human understanding, videos without captions and audio, is difficult to be properly understood by the model. We hope the release of MMVText can promote efficient video text reading, further enhancing automatic video description.

Video Text Automatic Translation. Another practical application is video text automatic translation, as shown in Figure. 1 (b). The application may be unnecessary for several professional teams or classic movies due to the professional translator or huge cost investment. But for international video-

sharing websites [1] [2] with millions of users, it isn't easy to apply multilingual caption and scene text in billions of videos. Therefore, efficient translation concerning caption text (e.g., overlap, song title, logos) and scene text (e.g., street signs, business signs, words on shirt) still need further exploration and research. The large-scale and multilingual MMVText contributes various real scenarios for the development of video text automatic translation.

Video Retrieval. Video retrieval with textual cues [5, 8] is also a very important application direction for video-and-text research, as shown in Figure. 1 (c). To the best of my knowledge, video retrieval with text information in the video is still almost a blank field of study and immature application in the industry. The most existing video retrieval methods are stiff combinations of text detection and recognition, invalid for the example with a sentence query. Besides, similar to video understanding, for the query of the sample video, *i.e., "The lakers play host to Golden State"*, we can not obtain the correct related video without scene text or caption information. The missing information needs to recover by understanding the video with key video text information such as *"GOLDEN STATE WARRIORS, LOS ANGELES LAKERS"*. The proposed MMVText with various text types (*e.g.,* caption, song title, logos, street signs, business signs) and annotation can promote the research concerning efficient video retrieval.

A.2.1 Limitations

Although the proposed MMVText supports all video text spotting tasks, *i.e., text detection, recognition, tracking end to end video text spotting*, the potential contributions for other tasks still need mining. For example, as shown in Figure. 1 (c), we do not provide the corresponding annotation (*i.e.,* the query for each video) and metrics concerning video retrieval, but the annotation and metric are easy to obtain due to text spotting annotation already existed. Therefore, there are still many potential contributions for other tasks on MMVText, we want to take these as the future research directions and provide a complete solution method.

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[2] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixe. Cvpr19 tracking and detection challenge: How crowded can it get? *arXiv preprint arXiv:1906.04567*, 2019.

[3] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *IEEE International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

[4] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

[5] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3040–3047, 2013.

[6] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Roadtext-1k: Text detection & recognition dataset for driving videos. In *IEEE International Conference on Robotics and Automation*, pages 11074–11080, 2020.

[7] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Workshops of European conference on computer vision*, pages 17–35, 2016.

[1] https://www.youtube.com/

[2] https://www.kuaishou.com/en

[8] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. *arXiv preprint arXiv:2104.01552*, 2021.

[9] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.

[10] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.