Commentary: Text From Corners: A Novel Approach to Detect Text and Caption in Videos

1. Introduction

The author aims to develop a robust and efficient method of automated detection of text in videos and static images with high accuracy.

This is an important goal as it allows us to effortlessly categorise, archive and retrieve desired video and image content containing text from a large dataset without the need to manually tag new content. As the amount of image data available to us continues to grow at an ever increasing rate this method of text detection would allow us to harness the information it provides without time wasted looking through unrelated content or manually categorising content, both of which are quickly becoming infeasible to do accurately with the scale of data produced [1].

While methods of text detection in video data do currently exist they are lacking in key areas. The sheer scale of content produced means that computational efficiency is at a premium and renders many otherwise effective methods impractical for the task proposed [2]. Likewise the scale of image data produced for many different purposes, using different equipment and in different locations results in high levels of variability in the images and text they depict. This means that many contemporary methods rely on assumptions in the consistent representation of the text or the image itself that do not hold, particularly across multiple languages or moving text [3]. These methods cannot be generalised for the task at hand.

Thus the benefit of the proposed method becomes clear. Computer vision experts tasked with categorising image data by text would no longer need to spend valuable time manually categorising content, building fragile automated models for their specific dataset or waiting on inefficient pre-existing algorithms.

In turn this ability to categorise a far larger amount of visual content than ever before would have huge benefits in many fields such as allowing biologists to easily access images of a desired species or aiding modern historians in amounting first-hand sources on a specific event based on text stamps of time and location. As a whole it allows society to more easily find the video and image content they desire; improved information sharing can only aid in making new discoveries and improving our quality of life overall.

2. Methods

Corner points are first extracted from the image to be used as the key feature in discerning text from non-text regions. This is an effective use of domain knowledge as corners are generally more orderly in text than non-text regions, making them a good discriminator. Corner points are also a stable and robust feature, less easily affected by noise than other commonly features such as edges.
The corner points are extracted from a frame using the Harris Corner Detector, an industry standard interest operator. It is a particularly adept choice in this use case due to its invariance to noise and rigid transformations. The drawback of the Harris Corner Detector that the authors neglect is the threshold hyperparameter. The robustness of this approach could then likely be improved with adaptive thresholding that has been used prior to great effect [4].

A binary mask of the corner points is then taken and dilated to merge nearby points into groups. Subsequently, morphological features on each group are taken to be later used to discriminate text from false alarms. The features are well chosen for these purposes as they utilise many aspects of how text is commonly displayed regardless of language or culture to build a robust feature set. The only caveat is that dilation is used without reference to size or shape of the structuring element. This omission is important as a rigid approach may mean the method would not generalise well over images of different sizes.

To account for moving text, dense motion estimation is performed across frames using the Lucas-Kanade approach. This method is suited to the task due to it's efficiency and it's assumption that a region will follow the same motion vector is likely to hold true as text very often moves with a constant velocity. If this assumption failed then I would suggest Particle Filtering as a usable, if less efficient, alternative.

The final classification of a possible text region is performed with a CART decision tree on the collected features of a single video, including new motion features. Decision trees are a efficient, non-parametric, non-linear classifier making them well suited as a robust binary classifier. Their usual drawback of high variance is counter-acted with the use of 10-fold cross validation. However a neural network would likely be a more effective classifier in this instance as it's accuracy would scale better with the large quantities of data available. The slower training and loss of interpretability would likely be worth the increased accuracy in this use case.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

3. Results

The text detection method is first evaluated using only videos with static text. Across 842 video shots and 7578 image frames the method was shown to perform extremely well, with 94.77% recall across videos and 86.48% recall and 93.24% precision across individual frames.

The success of the method is particularly salient when compared to a texture-based model. The proposed model records similar results (+1.89% precision, -6.62% recall) and was shown to be 15 times more computationally efficient (0.25s vs 3.8s). This massive increase in efficiency while retaining performance on par with an industry standard approach will likely be enough for many to adopt the proposed method. However this comparison is done over a subset of only 500 frames, drawing the results into question. An experiment using a more complete dataset which showed similar results would likely improved credibility of the proposed method for many potential users.

To evaluate moving caption detection a new dataset of 774 videos containing both moving and static text is used, although half was set aside for training the decision tree. On this dataset the method appears to perform very well, with an accuracy of over 90% on both video types.

However this evaluation method is lacking in many respects. Firstly the evaluation set contains only 23 videos of moving text in total. While this ratio may be representative of how often text is static or moving in real applications such a small number of moving text videos is inadequate to evaluate accuracy with any level of assurance and is unlikely to persuade anyone in the field to adopt the approach.

The evaluation of moving text detection also no longer shows the precision and recall per frame as was given for static text detection prior. It is perplexing why these standard metrics were not provided and could possibly obfuscate major flaws with the methods ability to detect moving text over the course of a whole video. A further experiment on a larger evaluation set of moving text as well as providing frame level metrics of the method's results would no doubt be much more likely to convince potential users to adopt the proposed method.

Evaluation of both moving and static text detection was performed on content from movies and television in curated competition dataset. This provided in the appearance of the text and video itself and thus likely increases the viability of the results. However it should be noted that professionally captured video likely shares common trends that may not be present in amateur recordings, such as the quality and stability of the recording. One trend pointed out by the authors is that all recordings shared a uniform aspect ratio, something that obviously will not occur across all datasets. Therefore an experiment using a wider range of videos, including different aspect ratios and varying image quality to showcase the robustness of the proposed method would likely be necessary to see further adoption.

4. Conclusions

The paper presents a very well considered method for text detection in video content. Domain knowledge has been expertly utilised to bring together a range of computer vision techniques with robust and efficient performance for the task.

I find the key feature of corner points was a keen choice as a novel and powerful discriminating feature of text-rich regions. The morphological features of these regions chosen to rule out false alarms show a robust understanding of how text is displayed in video across contexts. The Lucas-Kanade approach for estimating text motion is insightfully utilised in a case where it's assumptions are met without compromise. The use of a decision tree, a provenly robust 2-class classifier, further shows a keen understanding of the field, especially as it's primary downsides are deftly mitigated with 10-fold cross validation. All of these techniques were chosen with a clear consideration for the efficiency required to meet their initial aims. This results in a method that does not sacrifice accuracy compared to pre-existing methods but is far more capable of keeping up with the increasingly large scale video datasets.

The paper however falters in its evaluation methods. The results of the evaluation metrics used are very promising for both static and moving text detection but they fail to adequately prove the proposed method is as general as the introduction claims is the aim.

The issues of "flexibility" and "robustness" the proposed model seeks to address simply cannot be considered solved with such small evaluation sets. Only 500 frames of static text being used to compare to the standard method and only 23 videos of moving text with no model for comparison. This is only compounded by the fact these datasets share many features such as a fixed aspect ratio.

In fact the weakness of these evaluation metrics draws the effectiveness of the moving text detection into question. With such a small sample size and no frame-level data or industry standard metrics like precision and recall little can be concluded about even this most general aim of the paper.

Therefore my recommendation for future research would be to further evaluate the proposed method to corroborate the claims made with a breadth of evidence that this paper did not provide. The claims of the model's "flexibility" and "robustness" should be tested by evaluating on a much wider variety of video content from a range of sources, with particular care taken to evaluate the detection of moving text using standard metrics. While the framework of the proposed method has no major detracting flaws it is also worth tweaking the individual techniques used in future research to optimise the approach. In particular alternatives to parametric methods like Harris Corner Detector and dilation as well as other classification methods could be used to balance between accuracy and efficiency and meet the needs of future applications.

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

References

[1] A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. arXiv preprint arXiv:1601.07140, 2016.

[2] X. Tang, X. Gao, J. Liu, and H. Zhang, “A spatial-temporal approach for video caption detection and recognition,” IEEE Trans. Neural Netw., vol. 13, no. 4, pp. 961–971, Jul. 2002.

[3] W. Kim and C. Kim, “A new approach for overlay text detection and extraction from complex video scene,” IEEE Trans. Image Process., vol. 18, no. 2, pp. 401–411, Feb. 2009.

[4] Vino G., Sappa A.D. (2013) Revisiting Harris Corner Detector Algorithm: A Gradual Thresholding Approach. In: Kamel M., Campilho A. (eds) Image Analysis and Recognition. ICIAR 2013. Lecture Notes in Computer Science, vol 7950. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39094-4_40

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder