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

Student Name Anonymous
Student zID Anonymous

1. Introduction

Video technology has been prevalent in society for many generations and is one of the most popular forms of media throughout the internet and television. The purpose of video is not only limited to social media or entertainment purposes but serves as a highly important artefact for security, context analysis, military intelligence and various other circumstances where retrieval of information is necessary. Due to this, the amount of multimedia data stored and available for analysis has largely increased and continues to further develop. Often the analysis of these videos, in particular the text and captions found within them, is required for retrieval, indexing or identification which means accurate and efficient methods must be utilised. The requirement for accuracy and efficiency, coupled with the copious amount of data available for analysis highlights the need for an automated system to detect text and caption for further processing of information.

The need for text detection is clearly evident however, there are various obstacles and important considerations as outlined in this paper and other peer-reviewed literature. Complex backgrounds, variations in video quality and contrast, motion blur, perspectives and angles as well as variations in texts all occlude the detection process which means various methods need to be placed to combat the issues and output a highly accurate result. The authors explain that there are various approaches such as a texture-based methods, connected component-based methods and edge-based methods. They further suggest that texture-based methods are limited as they are not suitable for large databases due to their complexity, connected component-based methods are limited when texts contain noise, different colours or textures and an edge based approach is limited in reliability when backgrounds contain a similar distribution to the edge strength. The authors explain that their unique approach uses corner points detection to identify the text region (regardless of language) with distinct features as well as detecting moving captions located in videos.

The benefit of solving this issue results in the automation of video data analysis alleviating humans from individually detecting information. This can be useful in military applications to increase data retrieval speed and efficiency, government applications such as legibility of license plates or even private business applications such as removing videos from their platform if the algorithm detects censored language. An advanced algorithm will make an improvement in the current field by increasing accuracy and speed in multiple varieties of datasets.

2. Methods

The formalized methods proposed in this paper are divided into two sections; *features for text detection* which involves extracting corner points for feature description and *moving caption detection* which uses optical flow for motion features and decision tree for caption classification. I think the authors set a strong foundation in separating the challenge like such as it encompasses the various error domains and builds a specific model to improve overall accuracy. Maintaining a precise corner point extraction and identifying each feature is crucial for a successful outcome. As explored in the section below, I do comment on the strengths and weaknesses of each specific techniques as some techniques used can be challenged and improved.

Harris corner detection is used to identify and extract the corner values from the image sequences of the video. A corner is defined as the precise location where two dominant and different edge directions meet. The paper identifies that a corner detection technique is superior to a mere edge detection approach due to more confident features for pattern representation. The Harris corner detection approach is further implemented as a result of minimalistic changes in accuracy in response to rotation, scale, noise and lighting alterations. Other methods for corner detection exists like contour based methods of first identifying the edges using a canny edge detector [1] and then analyse the different contour properties to determine the location of the corners. This method could lead to many false negatives and not detect all the corners which is a large disadvantage of the method. Model based methods to identify corners [2] also exist which use predefined models to fit different subsets within the image but these methods are less efficient than the method provided in the paper even after using machine learning implementations to improve the performance. I think that the Harris corner detector is overall effective in identifying the corners to a high level of precision.

Feature description is important in order to distinguish the corner features that correspond to the text section of the image. The authors first apply an image morphology dilation in order to remove isolated corner values which are assumed to not be a part of the text region. This assumption is dangerous as it realises highly on a low amount of error with the corner detector algorithm and also that texts are uniform throughout different regions of the image. Different region properties are used to identify features such as area, saturation, orientation, aspect ratio and position. Filtering features based on area pose some limitations as it assumes that a location of a few words (names in the corner of the

video, etc) isolated from the main text may not be text and thus removed. Saturation is a good filter to use as often the saturation value of the pixels in the bounding box differ to the rest of the image but again is not always the case. Orientation assumes the length of the text is consistent and not random which poses a limitation when analysing texts with videos that are not just listed at the bottom of the screen. Aspect ratio could remove a lot of positive values if the thresh ratio is not calibrated to a high degree hence could be detrimental to the process. Position assumes the text will most likely be at the bottom of the image which is not always the case. I think combining these features could work contingent that correct thresh ratios are utilised.

Optical flow was used to detect motion features by implementing the Lucas-Kanade algorithm. I think that using this algorithm is appropriate as well as extracting features every 5 frames to preserve spatial-temporal information. The authors also combine the motion features quite well by incorporating whether the pixel is in the frame or not as well as a motion vector.

Decision tree is used to classify the caption based on the information of the moving text features. I think that the authors chose the right machine learning technique to classify the problem, however, decision trees contain many hyperparameters that could either increase the complexity of the problem or improve it. I would have liked to see more information on maximum depth size or minimum leaf size and also the criterion for the quality of the split.

3. Results

A dataset from Star Challenge containing various multimedia sources ranging from movie segments, television news and other videos are used to test the text detection system. The authors test both dynamic and static situations not only in image sequences also in video shots which is used to test moving caption detection due to its dynamic use case. The test scenarios give a wide range of text and caption cases found at the start and end of a movie. What it lacks are more complex texts that are a part of the video and not imported post production. It would have been interesting to see how the algorithm performs detecting words situated in the video like words on buildings, t-shirt, etc.

The text detection testing on static images was conducted on a category called introductory captions which contained 842 video shots and 7578 images sequences. The authors used recall and precision matrix in order to evaluate the performance of the system. This is a standard quantitative evaluation method which provides a good indication of the precision of the algorithm. Their method detected 798 shots and missed 44 on the video level and detected 4289 frames and 625 missed with 290 false positives. In general, this evaluation is informative however there is a limitation as the video shots yielded a 0 false positive as all contained either text or caption. This means that not enough information and evidence is provided. The authors should have tested on images that did not contain any texts or caption and determine whether their algorithm would detect a false caption. It would be interesting because the nature of removing individual false positives relies on information like the area of the location that most likely contains a caption. This could mean that if various corners are identified, even if there would be no real caption, the program may detect an area that appears to have text in comparison to the other features. SVM approach is also utilized in testing the results show the authors approach yielded a precision of 93.24% compared to the texture based approach of 91.35%, however, an 86.48% recall compared to a texture based of 93.1%. The time cost, on the other hand, was a lot better with the authors approach (0.25sec./frame) compared to the texture based approach (3.8sec./frame). The moving caption detection tests were performed similarly using quantitative evaluation metric showing detection ratio of over 90%. The results did yield a high miss detection ratio of 8.7% of moving text shots compared to 4.6% of missed when it wasn't moving.

As a result of this, I think that users would like to obtain more information on false detection before determining whether to adopt the proposed method. Being able to yield good results and winning the challenge does prove the algorithm to be promising in some areas, however, in areas where precision is crucial, more evidence needs to be provided. Testing on various other datasets with a combination of text, captions, natural texts in video and instances where no text is evident needs to be considered.

4. Conclusions

To summaries the paper, the authors produced an interesting and quite effective method of automating the detection of text and caption in videos. The paper identifies the common issues faced when detecting text and formulates various methods to combat them. The authors use of the Harris corner detection technique proves to be effective and precise in identifying the corner values in each image sequence. The discriminative features for identification of text are productive in distinguishing between text and non-text areas, however, decreases effectiveness if no text is evident on the image. The use of optical based motion coupled with text features to detect moving text is good, however, I would advise to look at further optimising the decision tree. The algorithm works quite well when data is sure to contain captions or texts in a uniform location on the screen. The limitation of the system is evident when detecting words that are integrated within the context of the video. Solving this issue will prove useful in military and governmental applications and not just analysis of inputted text on a video. A future recommendation would be to look at incorporating features to solve issues such as lighting, perspective and background can affect text in a video. Another recommendation for further research is to extend the algorithm in order to recognise words and phrases to either store in a database or alert when found. Detecting text in another language and translating it in English would be another interesting extension to the paper.

References

[1] Canny J, “A computational approach to edge detection,” IEEE Trans. Pattern Anal. Mach. Intell., 8(6):679-698, 1986.

[2]Guiducci, Antonio. "Corner characterization by differential geometry techniques." Pattern Recognition Letters, 8(5):311-318, 1988.