An end to end system for subtitle text extraction from movie

Hossam Elshahaby *

Cairo University, Egypt

Abstract

A new technique for text detection inside a complex graphical background, its extraction, and enhancement to be easily recognized using the optical character recognition (OCR). The technique uses a deep neural network for feature extraction and classifying the text as containing text or not. An Error Handling and Correction (EHC) technique is used to resolve classification errors. A Multiple Frame Integration (MFI) algorithm is introduced to extract the graphical text from its background. Text enhancement is done by adjusting the contrast, minimize noise, and increasing the pixels resolution. A standalone software Component-Off-The- Shelf (COTS) is used to recognize the text characters and qualify the system performance. Generalization for multilingual text is done with the proposed solution.

Keywords: Error Handling and Correction, Multiple Frame Integration, grayscale text blocks.

Accepted on October 25, 2021

Introduction

A newly created dataset containing videos with different languages is collected for this purpose to be used as a benchmark. A new HMVGG16 Convolutional Neural Network (CNN) is used for frame classification as text containing or non-text containing, has accuracy equals to 98%. The introduced system weighted average caption extraction accuracy equals to 96.15%. The Correctly Detected Characters (CDC) average recognition accuracy using the Abbyy SDK OCR engine equals 97.75%

Image- and video-based multimedia information has become more essential in the sectors of information sharing and services in recent years. Content-based retrieval is an essential way for managing and searching huge amounts of multimedia data [1]. The right recognition of text from pictures and video will establish a firm foundation for getting the proper retrieval result in the field of content-based multimedia retrieval. As a result, learning how to extract text from a complicated background is critical for comprehending and retrieving photos and videos. Text extraction is usually divided into two parts: text region recognition and text segmentation. There are four types of text region detection methods: edge-based detection, texture-based detection, linked region-based detection, and machine learning detection. The edge detection method uses the edge detection operator to identify the image's edges, then filters or aggregates the possible text regions, and lastly filters out text regions using heuristic methods. This approach, despite its tremendous efficiency, has poor resilience when subjected to Complicated background noise.

Using picture texture characteristics, the texture-based technique determines if pixel points or pixel blocks correspond to the text. Although this approach can successfully recognise character regions in complicated backgrounds, it has a low operating efficiency since it must cope with the differential operation of the entire image. The approach based on linked area recovers the same colour text from the backdrop using picture segmentation or colour clustering algorithms. The idea behind this method is that all of the characters are the same colour. When the image's complex background regions have the same colour as the text, however, the test results are unsatisfactory. By developing a learning mechanism, a machine learning-based technique identifies text and nontext fragments. Because such techniques rely on sample selection to train the learning machine for classification, the similarity between training and test sample sets is insufficient to provide optimum detection results. Because the identified text region has a complicated backdrop, the text must be separated from it before it can be used in other programmes. The most common text segmentation methods rely on text colour and partial space information, and they can be divided into three categories: threshold methods, unsupervised clustering methods, and statistical model methods [8, 9], with the above methods only applying to grayscale text blocks with simple backgrounds. When the backdrop has the same or comparable colour component as the text, misclassification can occur, and when determining the number of kernel functions in a statistical model is challenging, the text region extraction from a complex video image does not perform well.

Text is placed immediately on top of the image in complicated backdrop video images. As a result, recognised text blocks are likely to have some unpredictably complicated picture backgrounds, which will obstruct segmentation. The colour and texture diversity of a complex background makes estimating the text colour challenging. Meanwhile, the backdrop of segmented text blocks only preserves a tiny portion of the original background picture, limiting the amount of information available owing to the fragmented texture, which is difficult to characterise using model creation. This research provides a text extraction approach for complicated video scenes based on the aforesaid analyses. The suggested technique consists of two main steps: text region coarse detection using multiframe corner matching and heuristic criteria, and video text region extraction using texture and SVM to precisely place the text region against a complicated backdrop. Multiframe corner matching is mostly used to resolve Harris corner filtration concerns in complicated backdrop video scenes. The approach is based on the relationship between consecutive multiframe pictures, with the help of the video text's temporal redundancy, and multiframe fusion to increase text identification accuracy. Heuristic criteria may be used to filter candidate text areas, and they can alter depending on the kind of scene, which can improve the algorithm's efficiency and minimise the false alarm rate to some amount. The local texture of the picture is described using the LBP histogram, the similarity tolerance between images is determined using SVM, and ultimately text in a complicated video scene is properly recovered

References

1. W. Wu, X. Chen, and J. Yang, “Detection of text on road signs from video,” IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 4, pp. 378–390, 2005.
2. K. Jung, K. I. Kim, and A. K. Jain, “Text information extraction in images and video: a survey,” Pattern Recognition, vol. 37, no. 5, pp. 977–997, 2004.
3. P. Dubey, “Edge based text detection for multi-purpose application,” in Proceedings of the 8th International Conference of Signal Processing, vol. 4, Beijing, China, November 2006.
4. Ar and M. E. Karsligli, “Text region detection in digital documents images using textural features,” in Proceedings of the International Conference on Computer Analysis of Images and Patterns, Vienna, Austria, August 2007. W. Kim and C. Kim, “A new approach for overlay text detection and extraction from complex video scene,” IEEE Transactions on Image Processing, vol. 18, no. 2, pp. 401–411, 2009.
5. C. S. Shin, K. I. Kim, M. H. Park, and H. J. Kim, “Support vector machine-based text detection in digital video,” in Proceedings of the 10th IEEE Workshop on Neural Network for Signal Processing (NNSP '00), pp. 634–641, Sydney, Australia, December 2000.

***Correspondence to:**

Hossam Elshahaby
Assistant Professor
Cairo University
Egypt