Text Extraction from Video and Natural Scene Image: A Brief Review

Shamnasana V[1], Binoy D L[2]

[1]MEA Engineering College, State Highway 39, Nellikkunnu – Vengoor, Perinthalmanna, Malappuram, Kerala, India

[2]MEA Engineering College, State Highway 39, Nellikkunnu – Vengoor, Perinthalmanna, Malappuram, Kerala, India

Abstract: ***The growth of multimedia made images and videos important source of information. Therefore, research in multimedia is increasingly focused on retrieval techniques. Text is a prominent source of information in video. Extracting text from videos indicate a challenging work. This paper provides a brief review on extraction of word text from videos and scene images. Although video text extraction techniques have been addressed in previous surveys, they were treated in either an image based framework or separated in tracking or enhancement sections. This review comprehensively surveyed a large number of techniques for extracting text information from images and videos.***

Keywords: CNN, OCR, SWT, SVM, Caption text, HMM

1. Introduction

The growth of smart phones and online social media provides large amounts of visual data. For example, YouTube[1]streamed approximately 100 hours of video per minute worldwide in 2014. Text has received increasing attention as a key and direct information sources in video. Text in video is divided into two. Caption text and scene text. Caption text is also known as graphic text or artificial text. Caption text provides good directivity and a high-level overview of the semantic information in captions, subtitles and annotations of the video, while scene text is part of the camera images and is naturally embedded within objects (e.g., trademarks, signboards and buildings) in scenes. Caption text is further divided into two subcategories: layered caption text and embedded caption text. Layered caption text is always printed on a specifically designed background layer, while embedded caption text is overlaid and embedded on the frame.

A wide variety of methods have been proposed to extract text from images and videos. Many video text extraction methods detect and recognize text in each sampled individual frame (i.e., frame by frame) without multiframe integration. It involves

- Text detection
- Text recognition
- Text tracking

Text detection is the task of localizing the text in each video frame. Text recognition means what the localized text reads. The goal of text tracking is to continuously determine the location of text across multiple dynamic video frames. There are a variety of recent challenges for text extraction in scene by robots and users, like heterogeneous background, varied text, non-uniform illumination, arbitrary motion and low contrast. Most previous video text detection methods are reviewed with local information within individual frames, with limited performance. There are a limited number of approaches for scene text detection in video, most of them focus on extracting text with local information. At the same time, there are a few techniques with spatial and temporal information utilization for detecting text within multiple frames.

Figure 1: Text tracking and detection results on sample video.

This paper presents a review of text detection, tracking and recognition methods and systems in video, with a special focus on recent technical advancements.

2. Literature Survey

Xu-Cheng Yin et al [1] categorized existing methods for text detection into three major groups: Connected component (CC) based methods, Region based methods (Sliding window based methods), Maximally Stable Extremal Regions (MSERs) with Stroke Width Transform (SWT) method. Connected component based methods extract character candidates from images by connected component analysis followed by grouping character candidates into text, probably with additional checks to remove false positives. This methods usually perform well for captions that have uniform color and regular spacing; however, this methods may not preserve the full shapes of characters due to color bleeding and the low contrast of text lines. Region based methods use a binary text/non-text classifier to search for possible text regions over windows of multiple scales and aspect ratios and then group the text regions into text. The classifier may utilize various features including color, edges, gradients, texture and other related region features to distinguish between text and background. Candidate text regions are first found with an edge map or gradient information in video frames. Subsequently, a refinement stage is conducted using heuristic rules or learned classifiers.

These methods are fast and overcome low contrast problems, but they produce many false positives when the background is complex. To overcome this problem, texture features are utilized to detect text in video frames. The base techniques of texture features use various methods (e.g., Gabor filter, fast Fourier transform, spatial variance, wavelet transform, or multi-channel processing) to calculate the texture of blocks. Then, proper classifiers are employed to classify text blocks and non-text blocks. These techniques fail when text-like textures appear in the background. Some other methods integrate hybrid features to distinguish text from the background such as texture and edge features, color and edge/gradient features, and wavelet and color features. The MSER based method automatically detect and recognize text in natural images. This is a common task performed on unstructured scenes. Images with undetermined or random scenarios are called unstructured scenes . For example, detection and recognition of text automatically from captured video to alert a driver about a road sign is possible. This is different from structured scenes, which have scenarios where the position of text is known beforehand.

Shashank Shetty et al [2] proposes a new method for text recognition. The task performed is divided into three step approach. It combines the text detection and text recognition from the video frame. The video frame creation includes dividing of the video into an individual frames. These individual frames are grabbed and passed to the remaining two phases. The text detection is a two-step approach. It involves text localization phase and text verification phase. The text recognition includes text verification phase and optical character recognition phase. The final result of this paper is the detection of the text from the video frames in a word file. Experimental results of the proposed approach shows the accuracy level of Optical character recognition (OCR) in terms of text extraction.

Video text recognition is conventionally performed using existing OCR techniques. Text regions are first segmented from video frames and then fed into an OCR engine. The recognition performance based heavily on text segmentation/binarization (removing the background) and may suffer from noise and distortion in complex videos.

Max Jaderberg et al [3] presented an end to end system for text spotting, localising and recognising text in natural scene images and text based image retrieval. This system uses a region proposal mechanism for detection and deep convolutional neural networks for recognition. It is based on a novel combination of complementary proposal generation techniques to ensure high recall, and a fast continuos filtering stage for improving precision. Convolutional neural networks are trained to perform word recognition on the whole proposal region. At that time, departing from the character classifier based systems of the past. These networks are trained only by the data generated by the synthetic text generation engine. It does not require manually tagged data. In recent years, mainstream segmentation-based word recognition techniques typically over-segment the word image into small segments, combine adjacent segments into candidate characters, classify them using CNNs or gradient feature based classifiers, and find an approximately optimal word recognition result using beam search , Hidden Markov Models , or dynamic programming . Word spotting methods usually calculate a similarity measure between the candidate word image and a query word. Impressively, some recent methods design an appropriate CNN architecture and train the CNNs directly on holistic word images or use label embedding techniques to enrich the relations between word images and text strings. CNNs were initially used for recognition of handwritten digits. They were then applied faithfully on many problems of pattern recognition. Haojin Yang et al [4] proposed a workflow based method for video text detection and recognition. In the text detection stage a fast localization-verification scheme is developed. An edge-based multi-scale text detector first identifies potential text candidates with high recall rate. Then, an image entropy-based filter is used to refine the detected candidate text lines. Finally, Stroke Width Transform (SWT) and Support Vector Machine (SVM)-based verification procedures are applied to eliminate the false alarms. For text recognition, a novel skeleton-based binarization method is developed in order to separate text from complex backgrounds to make it processible for standard OCR (Optical Character Recognition) software [5].

Yiqing Wang et al [6] introduce a video text detecting and tracking approach. The clear binary text images obtained can be processed by OCR (Optical Character Recognition) software directly. This approach includes two parts, first one is stroke-model based video text detection and localization method, the second is SURF (Speeded Up Robust Features) based text region tracking method. In detection and localization approach, model and morphological operation is used to roughly identify candidate text regions. Combine stroke-map and edge response to localize text lines in each candidate text regions. Several heuristics and SVM (Support Vector Machine) used to verifying text blocks. The core part of the text tracking method is fast approximate nearest-neighbour search algorithm for extracted SURF features. Text-ending frame is determined using SURF feature point numbers, while, text motion estimation is based on correct matches in adjacent frames.

Sangheeta Roy et al [7] proposed a method which combine Hidden Markov Model (HMM) and Convolutional Neural Network (CNN) to achieve good recognition rate. Sequential gradient features with HMM help to find character alignment of a word. The character alignments are verified by Convolutional Neural network (CNN)

3. Conclusion

This review presents a detailed survey of methods used for text extraction. The process of extracting text from video and natural scene images involves series of steps. Text detection, text recognition and text tracking, where text tracking, tracking based detection, and tracking based recognition are specifically summarized and highlighted. More importantly, major technological trends and directions are discussed in detail, with the intention of identifying the

open issues and potential directions for future research from the current research literature.

References

[1] A. Yin, Xu-Cheng, et al. "Text detection, tracking and recognition in video: a comprehensive survey." *IEEE Transactions on Image Processing* 25.6 (2016): 2752-2773.
[2] Shetty, Shashank, et al. "Ote-OCR based text recognition and extraction from video frames." *Intelligent Systems and Control (ISCO), 2014 IEEE 8th International Conference on*. IEEE, 2014.
[3] Jaderberg, Max, et al. "Reading text in the wild with convolutional neural networks." *International Journal of Computer Vision* 116.1 (2016): 1-20.
[4] Yang, Haojin, Bernhard Quehl, and Harald Sack. "A framework for improved video text detection and recognition." *Multimedia Tools and Applications* 69.1 (2014): 217-245.
[5] Tian, Shu, et al. "Scene text detection in video by learning locally and globally." *Proc. 25th Int. Joint Conf. Artif. Intell.(IJCAI)*. 2016.
[6] Yusufu, Tuoerhongjiang, Yiqing Wang, and Xiangzhong Fang. "A video text detection and tracking system." *Multimedia (ISM), 2013 IEEE International Symposium on*. IEEE, 2013.
[7] Roy, Sangheeta, et al. "New Tampered Features for Scene and Caption Text Classification in Video Frame." *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE, 2016.

Author Profile

Shamnasana V received her B.Tech degree in Information Technology from the MES College of Engineering, Kuttippurm, Kerala, in 2013. Right now she is pursuing her M. Tech degree in Computer Science and Engineering at MEA Engineering College, Kerala from 2015 to 2017. Her research interests lies in Image Processing.

Binoy D L is a post graduate in Computer Science and Engineering from Karunya University Coimbatore in 2012. He graduated from College of Engineering Adoor in 2009. He is currently working as Assistant Professor in MEA Engineering College, Kerala. His research interest lies in Image Processing.