A Heuristic Approach to Caption Enhancement for Effective Video OCR

Lei Xie\textsuperscript{1,2} and Xi Tan\textsuperscript{1,3}

\textsuperscript{1} School of Computer Science, Northwestern Polytechnical University, Xi’an, China
\textsuperscript{2} Human-Computer Communications Laboratory
The Chinese University of Hong Kong, Hong Kong SAR
\textsuperscript{3} Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen, China
lxie@nwpu.edu.cn

Abstract. We present a heuristic approach to enhancing speech synchronized captions for video OCR, as a pre-process for subsequent tasks of multimedia indexing, segmentation and retrieval. We use a bi-search based caption transition detection method to improve efficiency, which adopts a simple heuristics that the same caption content usually lasts for a period for stable viewing. We propose a combination of color mask, changing mask and region mask to perform caption enhancement based on the discriminative characteristics of captions and backgrounds. Elaborate enhancement on individual characters is further used to remove small background residues. OCR experiments show that our caption enhancement approach brings a high character accuracy of 89.24%.

Keywords: Video OCR, caption transition detection, caption enhancement, video retrieval, multimedia retrieval.

1 Introduction

With the proliferation of videos from TV and the Web, extraction of superimposed captions in videos (i.e. Video OCR) has drawn much interest for multimedia segmentation, categorization and retrieval \cite{123}. Video captions contain rich content information such as names, locations, dates and time. Speech-synchronized captions in news archives are valuable for efficient news indexing and retrieval. However, extracting caption texts in videos is not a trivial task. Mature OCR technologies for document images cannot be directly used to recognize video captions since caption characters are often embedded in complex and dynamic instead of clear and static backgrounds as in document images. Captions usually hold small areas with low resolution characters and lossy video compression further degrades the quality. Video segmentation and categorization tasks desire the need of extracting the timing information of captions.

Approaches in Video OCR have mainly focused on caption/text detection and enhancement \cite{456}. Caption detection locates the caption regions and finding out caption (dis)appearance frames (i.e. caption transition detection). Upon detection of captions, image enhancement is adopted to enlarge the discrimination between caption characters and backgrounds. Tang et. al. have proposed a
spatial-temporal approach for Chinese caption detection and enhancement [4], where fuzzy clustering neural network (FCNN) is introduced for caption transition detection; multi-frame integration methods (e.g. minimal pixel search and frame averaging) and morphological operations are adopted for caption enhancement. FCNN-based caption transition detection is effective but it induces frame-by-frame feature extraction and classification that result in a low efficiency. Minimal pixel search for caption enhancement may not necessarily lead to a high caption-background contrast when backgrounds and captions share a similar or nearly the same color, which degrades the OCR performance seriously.

In this paper, we present a heuristic approach to enhancing speech synchronized captions for effective Video OCR as a pre-process for subsequent tasks of news segmentation and retrieval. We use a bi-search based caption transition detection method that improves efficiency using a simple heuristics that the same caption text usually lasts for a bunch of frames. We propose three mask operations (i.e. color mask, changing mask and region mask) to perform caption enhancement based on the discriminative characteristics of captions and backgrounds. Elaborate enhancement on individual characters is further adopted to remove small background residues. Our approach outperforms Tang’s approach [4] tested on the videos from the same source.

2 System Overview

The block diagram of our system is shown in Fig. 1. First, the caption transition detection module filters out video frames without captions and detects transitions between caption texts. As a result, the video sequence is segmented into distinct image groups either sharing the same caption text or without captions. The image groups with captions are then fed into the caption-level enhancement module and the character-level enhancement module, resulting in an enhanced caption image. The caption-level enhancement module processes the entire caption area by the proposed mask operations, and the character-level enhancement module segments each caption into individual characters and removes noisy background residues. Finally, the OCR software accepts the caption image and generates text outputs. Our system only processes the caption region since its location is known as a priori for a particular TV program.

Fig. 1. The block diagram of our system