creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:48009-200000224447

Doctor of Philosophy

Design and Implementation of Automatic Person Indexing and Retrieval System Based on the Overlay Text in the News Interview Video Sequences

The Graduate School of the
University of Ulsan
Department of Electrical/Electronic
and Computer Engineering
Lee, Sanghee

Design and Implementation of Automatic Person Indexing and Retrieval System Based on the Overlay Text in the News Interview Video Sequences

Supervisor: Kang-Hyun Jo

A Dissertation

Submitted to
the Graduate School of the University of Ulsan
In Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Engineering

by

Lee, Sanghee

Department of Electrical/Electronic and Computer Engineering
Ulsan, Korea
August 2019

Design and Implementation of Automatic Person Indexing and Retrieval System Based on the Overlay Text in the News Interview Video Sequences

This certifies that the dissertation
of Sanghee Lee is approved.

Committee Chair Prof. Hee-Jun Kang

Committee Member Prof. Young-Soo Suh

Committee Member Prof. Jang-Sik Park

Committee Member Prof. Hyun-Deok Kang

Committee Member Prof. Kang-Hyun Jo

Department of Electrical/Electronic and Computer Engineering
Ulsan, Korea
August 2019

ACKNOWLEDGMENTS

I would like to say thank all those who accompanied and supported me during the doctor course.

First of all, I would like to take this opportunity to express my deepest gratitude to my supervisor Professor Kang-Hyun Jo, for allowing me the opportunity to be a part of his research group and for his brilliant insight, patient guidance and constant encouragement which help me to complete this work. It is very honored for me to study with him. Without his support, the completion of this dissertation would have never become possible, and have remained a dream.

I sincerely thank Professor Hee-Jun Kang, Professor Young-Soo Suh from the University of Ulsan, Professor Jang-Sik Park from the Kyungsung University, and Professor Huyn-Deok Kang from the Ulsan National Institute of Science and Technology (UNIST) who are other members of my committee. I would like to express my deepest gratitude for being gladly given time for my work and giving advice on the lack of my thesis and things I did not think about. And I would like to thank Professors at Graduate School of Electrical/Electronic and Computer Engineering, the University of Ulsan for their brilliant teaching.

I am grateful to my former and current group colleagues of our laboratory. They are a great researcher and very good friends. Especially I would like to say thank the Korean students, Dongwook, Youlkyeong, Giseok, Bumsuk, Jinsu, Hansung, Jungwon, and Byungseok for their sincerely help during studying.

I would like to thank my parents-in-law and their family, my younger sister and her family since I always received from their encouragement. Most of all, I would like to especially thank my wife Seung-Eun, my two sons Donggeun and Dongjin for their endless love, understanding, and encouragement.

Finally, I dedicated this dissertation to my mother in heaven. My mother was always my friend and guide in my life.

ABSTRACT

Design and Implementation of Automatic Person Indexing and Retrieval System Based on the Overlay Text in the News Interview Video Sequences

Graduate School of the University of Ulsan,
Department of Electrical/Electronic and Computer Engineering,
Ulsan, Korea

Lee, Sanghee

With the advent of the digital age, a vast amount of video data has been created by consumers and professionals over the last few decades. And the advances in the data capturing, storage, and communication technologies have made vast amounts of video data available to consumer and professional applications. The tremendous increase in the use of video data entails a need to develop effective methods to manage these multimedia resources by their content. In response to such demands, many researchers have been motivated to develop powerful indexing systems to ensure easy access to the relevant information, navigation, and organization in the vast repositories of video data.

Recognizing the overlay text embedded in images and videos provides high-level semantic clues which enhance tremendously the automatic image and video indexing. These texts contain a more concise and direct description of the content of the video. Therefore, the overlay text plays an important role in the automated content analysis systems such as the scene understanding, indexing, browsing, and retrieval.

Especially, the overlay text in the broadcasting news video sequences provides more meaningful of the content than any other type of videos. The detection and recognition of the overlay text have become a hot topic in news video analysis such as identification of person or place, name of the new-worthy event, date of the event, stock market, other news statistics, and news summaries.

This dissertation proposes a novel approach to extract meaningful content information from the broadcasted news video sequences by collaborative integration of image understanding and natural language processing. As an actual example, we developed a person browser system that associates faces and overlaid name texts in videos. This is given news videos as a knowledge source, then automatically extracts face and name text association as content information. The proposed framework consists of the text detection module, the face detection module, and the person indexing module.

For the preprocessing step, the proposed system makes the sub-clip based on the beginning frame for only focusing on the frames with overlay text. In the text detection module, the system executes overlay text detection and separates the name text line. And the system processes detection and extraction of the overlay text, and text recognition by optical character recognition (OCR). In the face detection module, the face thumbnail is extracted. The face detection module makes the representative thumbnail of the interviewee. And the person indexing module generates automatically the index metadata by named entity recognition (NER). And finally, the person indexing database is automatically made by combining the recognized text with the face thumbnail.

The successful results of person information extraction reveal that the proposed methodology of integrated use of image understanding techniques and natural language processing technique is headed in the right direction to achieve our goal of accessing real contents of multimedia information.

Contents

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

OPN	Overlay Person Name
OCR	Optical Character Recognition
NER	Named Entity Recognition
MPEG	Moving Picture Experts Group
HSV	Hue-Saturation-Value
HLS	Hue-Lightness-Saturation
RGB	Red-Green-Blue
LBP	Local Binary Pattern
HOG	Histogram of Gradient
DCT	Discrete Cosine Transform
MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
CAMSHIFT	Continuously Adaptive Mean-SHIFT
ASCII	American Standard Code for Information Interchange
ROI	Region of Interests
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
PER	Person name tag
ORG	Organization tag
LOC	Location tag
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative

Chapter 1

Introduction

1.1 Motivation and Background

With the advent of the digital age, a vast amount of video data has been created by consumers and professionals over the last few decades. And the advances in the data capturing, storage, and communication technologies have made vast amounts of video data available to consumer and professional applications. The tremendous increase in the use of video data entails a need to develop effective methods to manage these multimedia resources by their content. In response to such demands, many researchers have been motivated to develop powerful indexing systems to ensure easy access to the relevant information, navigation, and organization in the vast repositories of video data.

Multimedia entities carry two types of information: content-based information and concept-based information. The content-based information of low-level features, such as colors, shapes, edges, frequencies, the energy of the signal, could be automatically identified and extracted. However, the concept-based information consists of high-level

features, called also semantic features. This information indicates what is represented in the video or what is meant by the audio signal. In general, this information is hardly identified automatically [1].

One of the goals of the multimedia indexing community is to design a system able to producing a rich semantic description of any multimedia document. To parse, index or abstract massive amounts of data, various video content analysis schemes have been proposed to use one or a combination of image, audio, and textual information in the videos. Among the semantic features, the text embedded in the video frames is of particular interests. Because the texts present in the video can provide important supplemental information for indexing and retrieval. And, the texts provide more intuitive information for people to understand the meaning of video content. For example, superimposed texts in news videos annotate the names of people and places or describe objects. With the help of the extracted related to the news, the news videos can be segmented and cataloged more accurately. Therefore, it enables applications such as keyword-based video search, automatic video logging, and text-based video indexing [1-3].

1.2 Kinds of Text in Video

For the problem “What is text?”, the answer could be structured edges, a series of uniform color regions, a group of strokes, or a kind of texture. However, there are many objects in natural scenes, such as leaves, fences or windows, that have similar edges, strokes, or texture properties with text, making it difficult to design effective feature representation to discriminate text. A better assumption could be that “Text is a hybrid of edges, connected-components, strokes, and texture”. Based on this assumption, this dissertation proposed for text detection work [4].

There exist mainly two kinds of text in videos. One is the scene text and the other is the overlay text (called graphics text or caption in other papers) such as Fig. 1.1. The scene text naturally exists in the image being record in native environments.

On the other hand, the overlay text is graphically generated and artificially overlaid on the image by a human at the time of editing.

The scene text is found in the street signs, text on cars, writing on shirts in natural scenes. The appearance of the text is occasional and usually brings less related to video information. Also, the difference between the different scene text is very big. However, The overlay text is used to describe the content of the video or give additional information related to it. The examples of the overlay text include the subtitles in news videos, sports scores. Recognizing the overlay text embedded in images provides high-level semantic clues which enhance tremendously an automatic image and video indexing. These texts contain a more concise and direct description of the content of the video. Therefore, the overlay text plays an important role in the automated content analysis systems such as the scene understanding, indexing, browsing, and retrieval [5-7].

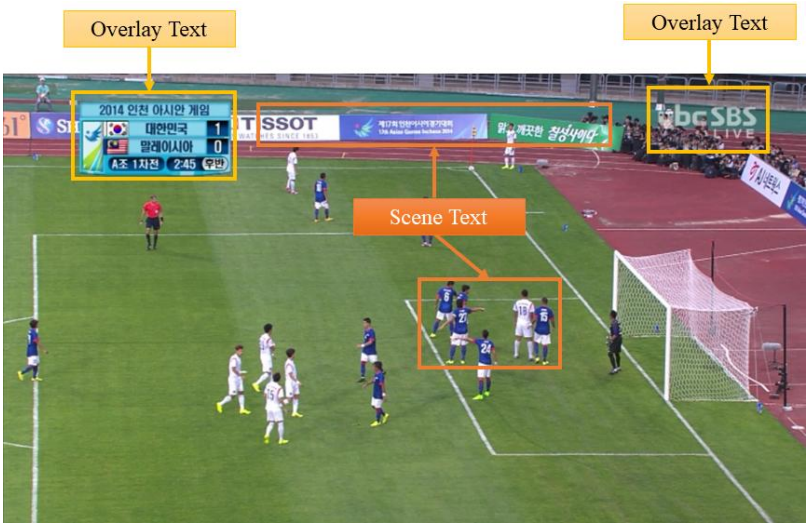


Fig. 1.1 The example of kinds of the text in video sequences

1.3 The Reason for Using News Videos

Especially, the overlay text in the news videos provides more meaningful of the content than any other type of videos. The detection and recognition of the overlay text have become a hot topic in news video analysis such as identification of person or place, name of the new-worthy event, date of the event, stock market, other news statistics, and news summaries [6-8].

Among these applications, the identification of the person from the overlay text raises a lot of interest in the information research community. The identification using the overlay person names (OPNs) has started to be investigated [2]. Since then the research area has raised a large amount of work, especially in face clustering tasks, face naming in captioned images, and recently, automatic naming within broadcasting videos [9-15].

Fig. 1.2 shows a typical composition of a news video. A news video consists of image sequences which may contain persons' face and transcripts which may contain persons' names. We are aiming at real content information extraction from news videos. News videos give us important content information, e.g., President... went to... to attend... meeting, Prime Minister... said... at that meeting, Senate leader ... talked about ..., etc. Looking at these types of content, it can be said that "who" information, i.e. the face and name association, is one of the most important information in news videos [16].

Fig. 1.2. The typical composition of news videos

Since the broadcasting video is produced by professional, many of the accepted production rules apply to the TV contents. For example, text and log are often overlaid onto the natural content in a structured manner, such as aligned text lines at the bottom or on the upper corners of the screen, to minimize the chance of covering the important content as shown in Fig. 1.3 [8].

And the common properties in most news video sequences are summarized as follows: The overlay text position is fixed, generally in the range of 1/3 from the bottom of the frame. The background of the text is usually opaque or translucent, and in most case, the color of the background is eye-catching, such as white, blue, yellow, and so on. Colors of text characters are often very distinguishable from the background color. The size and fonts of the overlay text from the same news video generally remain unchanged for a long term [8]. Therefore, this dissertation uses the above rule-based characteristics. And this rule-based classifier well distinguishes overlay text from any other scene in news video sequences.

Fig. 1.3. The example of the rule-based production in news interview video sequences

1.4 The Predominant Difficulties in Video OCR

Text appears in videos in a wide range of writings, fonts, styles, colors, sizes, orientation, and so on. Therefore, these problems make the exact modeling of all types of text almost impossible [17].

In natural environments, numerous man-made objects, such as buildings, symbols, and paintings, appear that have similar structures and appearances to text. Overlay text itself is typically laid complexity is that the surrounding scenery makes it difficult to discriminate text from non-text. Video compression and decompression procedures degrade the quality of overlay text in the video sequences. The typical influence of degradation is that they reduce characters sharpness and introduce touching characters. Therefore this effect makes basic task such as segmentation difficult [4].

Overlay text in the video sequences has different aspect ratios. To detect text, a search procedure with respect to location, scale, and length needs to be considered. As a result, high computational complexity is incurred. In the case of perspective distortion, text boundaries lose rectangular shapes and characters distort, decreasing the performance of recognition models trained on undistorted samples. Characters of italic and script fonts make it difficult to perform segmentation. Characters of various fonts have large within-class variations and form many pattern sub-spaces, making it difficult to perform accurate recognition when the characters class number is large [4].

Therefore, since the motivation behind the proposed method is semantic indexing, this dissertation concentrates on horizontally aligned overlay text in the broadcasting news. And in television news videos, the predominant difficulties in performing video OCR on overlay text are due to low-resolution characters and widely varying complex backgrounds. For the problem of complex backgrounds, an image enhancement method by multi-frame integration is employed using the enhanced resolution interpolation frames. Although complex backgrounds usually have movement, the position of video captions is relatively stable across frames. Furthermore, we assume that overlay texts have high-intensity values such as white pixels [2]. Therefore, this dissertation employs a technique to minimized the variation of the background using a time-based minimum pixel value search.

1.5 Dissertation Objective

Potential applications of the overlay text include the creation of the face-name association database, video annotation by person's name, and so on [16]. This dissertation proposes a novel approach to extract meaningful content information from the broadcasted news video sequences by collaborative integration of image understanding and natural language processing. As an actual example, we developed a person browser system that associates faces and names in videos. This is given news videos as a knowledge source, then automatically extracts face and name text association as content information.

Although humans seem to be able to do this easily, this process includes several complicated, and high-level processes: Text detection and recognition, Face detection and identification, and Named entity recognition. It is very hard for a computer to achieve these processes automatically. They are still far from complete, but it may be promising to properly integrate those techniques to get useful results. This dissertation takes the strategy to bring a face and name association to fruition. The proposed method is used production rule based name text line extraction and recognition, and dictionary-based named entity extraction, despite these being still incomplete.

When a description of all multimedia information is performed by a human, not only the tremendous amount of work is needed, but also it is difficult to maintain the consistency of human subjective thoughts. Therefore, the technology of automatic metadata generation and indexing is required to retrieve related information by data contents.

Until now, a common indexing method for various videos is impossible. Therefore, in order to handle video data efficiently, proper indexing technique considering characteristics of video data is needed. This dissertation aims at improving the convenience of news retrieval to users by efficiently and automatically indexing news that is broadcasted daily.

1.6 Dissertation Outline

This dissertation is presented an entire working pipeline for automatic person information extraction tailored specially for the broadcasted news interview video sequences as shown in Fig. 1.4. To do this goal, the proposed framework consists of the text detection module, the face detection module, and the person indexing database module as shown in Fig. 1.5.

For the preprocessing step, this system makes the sub-clip based on the beginning frame for only focusing on the frames with overlay text. In the text detection module, this system executes the overlay text detection and separates the name text line. And the system processes detection and extraction of the overlay text, and text recognition by optical character recognition (OCR). In the face detection module, the face thumbnail is extracted. The face detection module makes the representative thumbnail of the interviewee. And the person indexing database module generates automatically the index metadata by named entity recognition (NER). And finally, a person indexing database is automatically made by combining the recognized text with the face thumbnail.

The process of extraction carries detection, localization, tracking, extracting, enhancement and recognition of the text from a given image as shown in Fig. 1.4. Text detection refers to the identification of text in a given video frame. Text localization refers to determine the location of the text in the frame or sequence of the video frame. Text tracking is performed to cut back the interval for text localization and to maintain the position across adjacent frames. The precise location of text in a frame indicated by bounding boxes, the text still has to be divided from the background to use for its recognition. Text extraction is the stage that the text components are segmented from the background. Extracted text components are required enhancement because the text region usually has low-resolution and is susceptible to noise. The extracted text images are transformed into plain text using OCR technology.

The remainder of this dissertation is given as follows. Section 2 describes the relatives' works per step by step of the proposed system.

Fig. 1.4. The proposed entire working pipeline

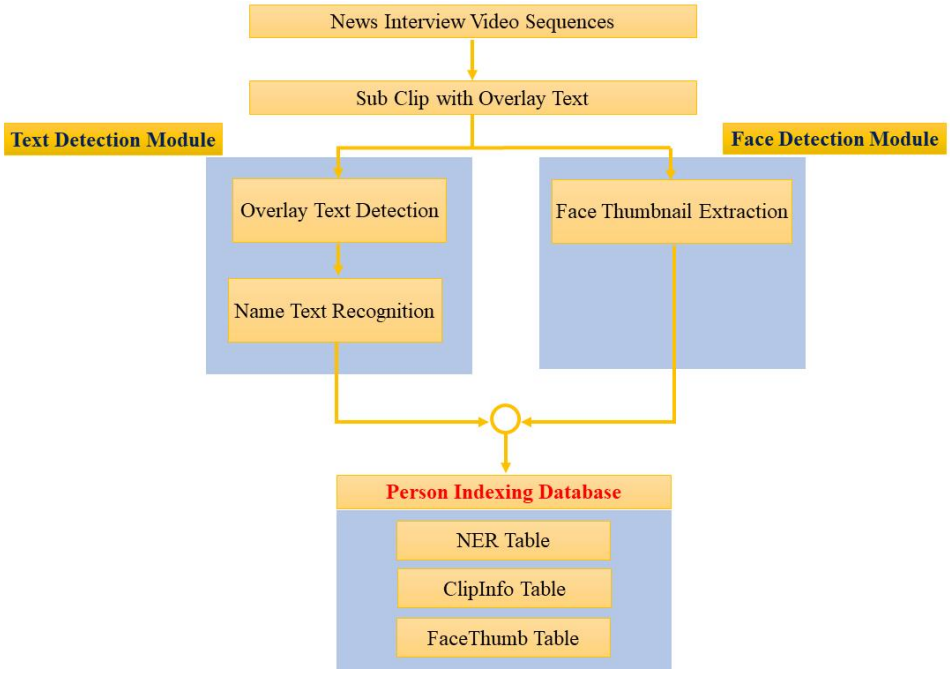


Fig. 1.5. The proposed framework for automatic person indexing

In section 3, for the preprocessing step, the proposed system makes the sub-clip based on the beginning frame for only focusing on the frames with overlay text. This dissertation is compared to the performance of the beginning frame identification by the experiment between the Canny edge detector and the Harris corner detector. Then, the appropriate processing method is proposed with the detection beginning frame with Canny edge.

Section 4 describes the name text line localization. This dissertation uses the temporal analysis of the broadcasted news videos to achieve a good accuracy of video text detection. And the method uses the rule-based characteristics in the production of a TV news program.

Section 5 describes personal information extraction. For the broadcasted interview video archives, this dissertation presents how to build automatically indexing by identifying the named entities in the extracted overlay name text.

Section 6 presents the person browser system using the proposed method and the experimental results.

In section 7, the conclusion is drawn and presents some ideas and suggestions for future works.

1.7 Unification of Words, Phrases, and Definitions in This Dissertation

In computer vision, several words, phrases, and definitions get a relative meaning. For instance, for text-based video indexing and retrieval, a typical system contains four main parts: Text Detection, Text Localization, Text Tracking, Text Extraction, and Text Recognition [18]. To easily discuss and compare our method and the others, we unify the commonly used words, phrases, and definitions as follows.

Text detection refers to the identification of text in a given frame. Text localization refers to determine the location of text in the frame or sequence of a video frame. Text tracking is performed to cut back the interval for text localization and to maintain the position across adjacent frames. The precise location of text in a frame indicated by bounding boxes, the text still has to be divided from the background to use for its recognition [19].

Text extraction is that the stage wherever the text components are segmented from the background. Extracted text components are required enhancement because the text region usually has low-resolution and is susceptible to noise. The extracted text images are transformed into plain text using OCR technology [19].

Chapter 2

Literature Review

2.1 Introduction

Text data present in multimedia business videos and images contain useful information for automatic annotation, indexing. In general, the process of extraction of information is detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image and video [19]. This chapter mainly analyzes the technologies of the process components.

2.2 Text Frame Detection

Text detection refers to the identification of text in a given video frame. Since not all the video frames contain text, it is a waste of time if detect and recognize every frame of the video, so a simple method is needed to detect the frames that contain the text.

Kim chosen frame form shots detected by a scene-change detection methodology as a candidate containing text. Kim's scene-change detection methodology isn't delineated thoroughly in his literature, he has mentioned that low threshold values are required for scene-change detection as a result of a very small region of images occupied by text regions. This approach is incredibly sensitive to scene-change detection. The text-frame choice is performed at an associated interval of two seconds for caption text within the detected scene frames. This can be an easy and economical resolution for video indexing applications that only need keywords from video clips, instead of the whole text [20].

Smith and Kanade proposed a method to detect the scene change based on the difference between two sequential frames and detect the text frame by the information of scene switch [21]. Gargi et al. performed text detection by assuming the amount of intra-coded blocks in P- and B- frames of a MPEG compressed video will increase, once a text caption appears [22].

Lim et al. made an easy assumption that text typically includes a higher intensity than the background. Author(s) counted the number of pixels that are less than a threshold and exhibited a big color distinction to their neighborhood. This methodology is very simple and fast [23]. Zhong Ji used a small window to scan the whole video frame, and extracted the mix characteristics from it, then extracted the text from the background by SVM classifier [24].

2.3 Text Localization

The objective of the text localization is to localize text components precisely as well as to group them into candidate text regions with as little background as possible. Since the text's color, size, font, and location are changeable, it is difficult to find a general method to separate it from the background. According to the features used and the ways they work, text location approaches can be divided into two categories: region-based approach and texture based approach [18, 25].

Region-based approach utilizes the different region properties between text and

background to extract text objects. This approach works in a bottom-up way by separating the image into small regions and then grouping character regions into text regions. Color features, edge features, and connected component methods are often used in this approach [25].

The texture-based approach uses distinct texture properties of text to extract text objects from the background. This approach works typically in a top-down way by extracting texture features of the image and then locating text regions. Spatial variance, Fourier transform, Wavelet transform, and machine learning methods are often used in this approach. Besides the above two categories, there are still some other text extraction approaches, such as hybrid based, compressed domain based, graphical model based, and so on [25].

2.3.1 Region-based Approaches

Text is often produced in a consistent and distinguishable color so that it contrasts with the background. Under this assumption, color features could be used to localize text in many papers. Color based text localization operates often simply and efficiently, although it is sensitive to multi color characters and uneven lighting, which can seriously degrade color features [20].

Jain et al. used color reduction to generate color layers, a clustering algorithm to obtain connected components and connected the connected components into text candidates with color similarity and component layout analysis [26]. Garcia et al. performed text extraction with a k-means clustering algorithm in the Hue-Saturation-Value (HSV) color space [27]. Karatzas et al. extracted text components with a split-and-merge strategy in the Hue-Lightness-Saturation (HLS) color space [28]. Chen et al. proposed using Gaussian mixture models in RGB, hue and intensity channels to localize text [29]. Nikolaou et al. show that the use of the mean shift algorithm to generate color layers could improve the robustness to complex backgrounds [30].

The family of edge or gradient-based approaches assumes that text exhibits a

strong and symmetric gradient against its background. Thus, those pixels with large and symmetric gradient values could be regarded as text components. Compared with color features, gradient or edge features are less sensitive to uneven lighting and multi-color characters. They are combined with such classifiers as artificial neural networks or AdaBoost to perform sliding window based text localization. However, they often have difficulty when discriminating text components with complex backgrounds having a strong gradient [20].

Wu et al. proposed using Gaussian derivative to extract horizontally aligned vertical edge, which is aggregated to produce chips corresponding to text strings if 'short path' exist between edge pairs [31]. Phan et al. proposed grouping horizontally aligned components of 'Gradient Vector Flow' into text candidates based on spatial constraints of sizes, positions, and color distances [32].

2.3.2 Texture-based Approaches

The main idea of texture based approaches is to consider that text and the background has a different texture. The texture-based approaches are more robust than region-based approaches, but its cost is bigger at the same time [18]. When characters are dense, the text could be considered as a texture. Texture features including Fourier Transform, Discrete Cosine Transform, Wavelet, Local Binary Pattern (LBP), and HOG have been used to localize text. Such features are usually combined with a multi-scale sliding window classification method to perform text localization. Texture features are effective for detecting dense characters, although they might not detect sparse character, i.e., signs in scene image which lack significant texture properties [20].

Kwang et al. used a small window to scan the whole image, then analyze its texture features by SVM to divide the center point of the window into text and non-text, then use Continuously Adaptive Mean-SHIFT(CAMSHIFT) to analyze text area and get text finally [33].

Chen et al. proposed a two-step method to detect and locate text in videos. The first step is to determine candidate text area: use Canny operator to get vertical

and horizontal edges of the image, then extend the edges in both vertical and horizontal directions by dilation operators. The second step is to determine the final location by four features: gray-scale spatial derivative features; distance map feature; DCT coefficients; constant gradient variance feature. Finally, multi-layer perceptrons (MLP) and SVM are used to classify the text objects [34].

Qian et al. used the method of Discrete Cosine Transform (DCT) [35]. Ji et al. used two texture features namely wavelet coefficients and Gray-level co-occurrence matrix for text detection along with SVM [36]. Peng et al. computed the feature using s-D Gabor filters and Harris corner detection [37].

Anthimopoulos et al. proposed a two-stage hybrid system for text detection in video frames. In the first stage text regions are detected based on the edge map of the image. An SVM classifier is used in the second stage, they used this classifier and a sliding window model to refine the result that got in the first stage. They pointed out that the most important aspect in designing the machine learning technique is the choice of features. These highly discriminating feature set based on a new texture operator which is called Local Binary Pattern(LBP) based operator [38].

2.4 Text Tracking and Extraction

The text localization step often introduces false positives because a small piece of components or patches may not contain sufficient information for classification. Ager text localization, holistic features of text regions are available for precise classification and verification. Due to many reasons, utility, enhancement speedup, and so on, tracking of the text in video has not been studied in great extends. Locating text in images, such as low resolution and complex backgrounds, these topics need to be more investigation [19].

2.4.1 Text Tracking

To enhance the system performance, it's necessary to consider temporal changes in a frame sequence. The text tracking stage will serve to verify the text localization results. Additionally, if text tracking may be performed in a shorter time than text detection and localization, this might speed up the general system. In cases wherever text is occluded in different frames, text tracking can be facilitated recover the original image.

Lienhart et al. represented a block-matching algorithm, that is an international standard for video compression like H.261 and MPEG, and used temporal text motion info to refine extracted text regions. The minimum mean absolute difference is used as the matching criterion. Each localized block is checked on whether or not its fill factor is higher than a given threshold value. For every block that meets the specified fill actor, a block-matching algorithm is performed [39].

Li et al. presented approach or many circumstances, including scrolling, captions, text printed on an athlete's jersey, etc. Authors used the sum of the square difference for a pure translational motion model, based on multi-resolution matching, to reduce the computational complexity. The text contours are used to stabilize for additional complex motions for the tracking process. Edge maps are generated using the Canny operator for the larger text block. Once a horizontal smearing method to cluster the text blocks, the new text position is extracted. However, since a translational model is used, this methodology isn't appropriate to handle scale, rotation, and perspective variations [40].

Swaki et al. projected techniques for adaptively acquiring templates of degraded characters in scene images involving the automated creation of content-based image templates from text line image [41].

Antani et al. [42] and Gargi et al. [43] utilize motion vectors in a compressed video MPEG-1 bit stream for tracking the text, based on the strategies of Nakajama et al. [44] and Pilu [45]. This methodology is implemented on the P and that I frames in MPEG-1 video streams. The original bounding box is then moved by the total of the motion vectors of all the macroblocks that correspond to the current bounding box.

2.4.2 Text Extraction

It is necessary for the text to be an enhancement and binarized processed before the process of OCR since the text area still complex background and low resolution.

SY. Liu et al. proposed a method to extract text from complex color structure background by color clustering in HSV space: first do the color clustering in HSV space, then separate te color layer (sub layer) which is get from the first step, put the text that have similar color and non-text into different layers, extract the text layer based on its features, finally combine the text layers with the same color [46].

L. N. Sun et al. proposed a method to do text binary processing based on Otsu's threshold and region filling [47].

Since most of the previous approaches to extracting label text from videos are based on low-level features, such as edge, color, and texture information. Kim proposed a novel framework to detect and extract the label text from the video scene. He found that there exist transient colors between inserted text and its adjacent background. So he firstly generated a transition map, and then a candidate regions. The projection of label text pixels in the transition map is used to localize the detected label text accurately and the text extraction is finally conducted [48].

Anoual et al. proposed a method: the first stage consisted of two steps: Edge detection and closed contour selection. And then they capitalized on the fact that a text area is a closed contour and used this texture information to discriminate text from non-text regions and verify regions in order to reduce false alarms and to generate valid text regions [49].

Shivakumara et al. proposed a segmentation scheme based on grayscale [50]. Wang Qi et al. used the gradient method to detect the edge of text, then mapped the text location extracted from the gradient image to the original video frame to get a large amount of accurate text pixels color. Use global mix Gaussian model to build a color model based on these color information. After modeling, use it to extract the text color layer from the subsequent text layer directly, and updating model parameters at the same time [51].

Shivakumara et al. proposed a method using Bayesian classification and boundary growing to detect multi-oriented scene text: they intersected the output of

Bayesian classifier with the Canny edge map of the input frame to obtain the text candidates, and a boundary growing method is introduced to traverse the multi-oriented scene text lines using text candidates [52].

2.5 Text-Based Video Indexing and Retrieval

Satoh et. al. proposes a novel approach to extract meaningful content information from video by collaborative integration of image understanding and natural language processing. As an actual example, authors integrate both visual (face, text caption) and audio (closed caption) to associate face and names in news video. This paper work on using video OCR for video indexing and retrieval. The success of their system demonstrates the benefits of a multi-modal approach for video indexing [16].

Lienhart presented the experimental results on video indexing using recognized text. After automatic text segmentation, the output is directly passed to a standard OCR software package to translate the segmented text into ASCII. This algorithm makes use of typical characteristics of text in videos in order to enable and enhance segmentation performance. Especially the inter-frame dependencies of the characters provide new possibilities for their refinement. Then, a straightforward indexing and retrieval scheme is introduced [53].

Li and Doermann present text-based video indexing and retrieval by expanding the semantics of query word and using the Glimpse to perform approximate matching instead of exact matching. To solve the problem that the text in digital video is usually of poor quality, this paper uses an approximate word matching algorithm instead of exact word matching. To challenge the problem that the text in digital video is usually very terse and may lack semantic breadth, this paper considers a limited class of video types, such as news, finance, or sports. Therefore, it is possible to build a local semantic dictionary. Aa result, this paper solve the expected OCR errors and the lack of semantic breadth in video text [54].

Jawahar, et al. propose an approach that enables search based on the textual information present in the video. Regions of textual information are identified within the

frames of the video. Video is then annotated with the textual content present in the images. Traditionally, OCRs are used to extract text within the video. The choice of OCRs bring in may constraints on the language and the font that they can take care of. Authors hence proposed an approach that enables matching at the image-level and thereby avoiding an OCR. Videos containing the query string are retrieved from a video database and stored based on the relevance [55].

Cees et al. propose an automatic video retrieval method based on high-level concept detectors. This paper aims to throw a bridge between the two fields by building a multimedia thesaurus, i.e., a set of machine-learned concept detectors that is enriched with semantic descriptions and semantic structure obtained from WordNet. Given a multimodal user query, this paper identifies three strategies to select a relevant detector from this thesaurus, namely: text matching, ontology querying, and semantic visual querying [56].

In the literature, many algorithms have been presented but any techniques didn't provide satisfactory performance.

2.6 Named Entity Recognition

Information extraction (IE) aims to locate inside a text passage domain-specific and pre-specified facts (for example, in a passage about athletics, facts about the athlete participating in a 100m event, such as his name, nationality, performance, as well as about the specific event, such as its name). Information extraction can be defined as the automatic identification of selected entities, relations or event in free text. Named entity recognition (NER), where entity mentions are recognized and classified proper types for the thematic domain such as persons, places, organizations, dates, and so on [57].

Named entity annotation for English is extensively examined in the literature and a number of automated annotation tools exist. The majority of them such as GATE, Stanford NLP tools, Illinois NLP tools, Apache OpenNLP library, LingPipe contain a number of reusable text processing toolkits for various computational problems such as tokenization, sentence segmentation, part of speech tagging, named entity extraction,

chunking, parsing, and frequently, co-reference resolution. GATE additionally comes with numerous reusable text processing components for many natural languages [57].

Another category of NER tools involves stand-alone automated linguistic annotation tools such as Callisto for task-specific annotation interfaces e.g., named entities, relations, time expression, and so on. MMAX2 which uses stand-off XML and annotation schemas for customization, and Knowtator which supports semi-automatic adjudication and the creation of a consensus annotation set [57].

2.6.1 Using Rule-based

One approach to performing NER is to use a combination of lists and regular expressions to identify named entities. This approach was popular in the early research on named-entity recognition systems but has become less popular because such a system is difficult to maintain for the following reasons: Maintaining the lists is labor intensive and inflexible. Moving to other languages or domains may involve repeating much of the work. Many proper nouns are also valid in other roles (such as Will or Hope). Said another way: dealing with ambiguity is hard. Many names are conjunctions of other names, such as the Scottish Exhibitions and Conference Center where it's not always clear where the name ends. Names of people and places are often the same-Washington (state, D.C., or George) or Cicero (the ancient philosopher, the town in New York, or some other place). It's difficult to model dependencies between names across a document using rules on regular expressions [58].

Rule-based approaches can perform nicely within specific well-understood domains, and shouldn't be discarded completely. For a domain such as capturing length measurements, where the items themselves are typically rare and the number of units of length bounded, this approach is probably applicable. Many useful resources are available publicly to help in bootstrapping such a process for a number of entity types. Basic rules and some general resources are available at the CIA World Fact Book and Wikipedia. Also available are dictionaries of proper nouns, along with domain-specific resources like the Internet Movie Database or domain-specific knowledge that can be

effectively utilized to achieve reasonable performance, while minimizing the work required [58].

2.6.2 Using Statistical Model-based

A less brittle approach that's easy to extend other domains and languages and that doesn't require creating large lists (gazetteers) to be maintained is much more desirable. This approach is to use a statistical classifier to identify named entities. Typically the classifier looks at each word in a sentence and decides whether it's the start of a named entity, the continuation of an already started entity, or not part of a name at all. The classifier needs to be trained on a collection of human-annotated text to learn how to identify names. Some of the advantages of such an approach include these: List can be incorporated as features and as such are only one source of information. Moving to other languages or domains may only involve minimal code changes. It's easier to model the context within a sentence and in a document. The classifier can be retained to incorporate additional text or other features [58].

The main disadvantage of such approaches is the need for human-annotated data. Whereas a programmer can write a set of rules and see them being applied to a collection of text immediately, the classifier will typically need to be trained on approximately 30,000 words to perform moderately well. Though annotation is tedious, it doesn't require the set of specialized skills needed for rule crafting and is a resource that can be extended and reused [58].

Chapter 3

Video Clip Segmentation

3.1 Introduction

Since a TV news program comprises huge numbers of frames, it is computationally prohibitive to detect each character in every frame. Therefore, to increase processing speed, we first roughly detect text regions in groups of frames. Some known constraints of text regions can reduce processing costs. A typical text region can be characterized as a horizontal rectangular structure sharp edges since characters usually form regions of high contrast against the background [2].

By observing a large quantity of TV news programs, as shown in Fig. 3.1, the appearance and disappearance of the overlay text occur suddenly or slowly in most news videos. Since the overlay texts appearing or disappearing in news videos are not changed faster than the scene content in a shot, the overlay text extraction from every frame is unnecessary and time-wasting. By precisely locating the critical frame where each overlay text appears or disappears, the overlay text detection can be focus solely on the frame contained the overlay texts.

Therefore, to achieve a good result of the overlay text detection in news video sequences, this dissertation proposes the identification of the overlay text beginning frame. The beginning frame is defined as the frame which has abruptly difference at the edge density of the current frame and previous frame and a little difference at edge density of the current frame and next frame. The proposed method acts as the second pass over the output of a text detector in the entire video sequences. And the proposed system makes the sub-clip based on the beginning frame for only focusing on the frames with overlay text.

This dissertation uses the density of the text edge obtained by the edge detector. If a frame has a high edge density, it has the overlay text. Otherwise, a frame not contains the overlay text. This dissertation is compared to the performance of the beginning frame identification by the experiment between the Canny edge detector and the Harris corner detector. Then, the appropriate processing method is proposed with the detection beginning frame with Canny edge.

3.2 Characteristics of Overlay Text in Video

By observing a large number of TV programs, this dissertation defines the characteristics of overlay text in the video sequences as follows that might be useful for a video indexing system. The observed characteristics of overlay text could serve as features for a text detection, localization, segmentation and recognition processing.

(a) frame 1 (b) frame 15 (c) frame 30

Fig. 3.1. The example of overlay text appearance in broadcasting news video sequences

Overlay texts are meant to be read by humans at a range of viewing distances in a limited time. Therefore, there is usually a minimum size of characters. On the other hands, the upper bound on character size is much looser. Overlay text can often be as large as half the frame height. Characters of various fonts have large within-class variations and form many pattern sub-spaces, making it difficult to perform accurate recognition when the character class number is larges [2, 22].

The characters in a single overlay text tend to have a similar color. Even if the color does vary across the overlay text, it varies in a gradual way so that adjacent characters or character segments have very similar colors [22].

The contrast between text and background is often is high. This is because the background over which text is composited is varying both spatially and temporally and it often happens that the color of characters is similar to that of the surrounding frame region. However, even if this occurs it is usually true only for a portion of the overlay text since otherwise, the compositor would have chosen a different color for the text [22].

Characters belonging to an overlay text string are usually horizontally aligned. This is a very strong feature. Also, the aspect ratio of the individual characters as well as that of the entire overlay text lies in a certain range [22].

Most overlay texts consist of a certain minimum and a maximum number of characters. In other words, at least one word but usually not more than a few. This is not true of single or double character logos or symbols, for example, the logo of the TV station, but since they are usually not relevant to the current scene content and hence not useful for a text indexing system, they can be ignored. Also, since overlay texts do not contain more than a few words, the characters are usually well separated. Therefore, touching characters, a thorny issue in document character recognition is not as important in this application [22].

Overlay texts tend to retain their size and font over multiple frames. However, this is not always true. The text often has special graphic effects added such as zooming up or down. The position of overlay text can change, such as scrolling text, but does so in a very uniform way, either vertically or horizontally. Successive frames exhibit very small jitter in overlay text position, usually not more than one or two pixels. Therefore, this is a very strong feature [22].

3.3 Comparison of the Beginning Frame Identification Methods

3.3.1 Method by the Canny Edge Detector

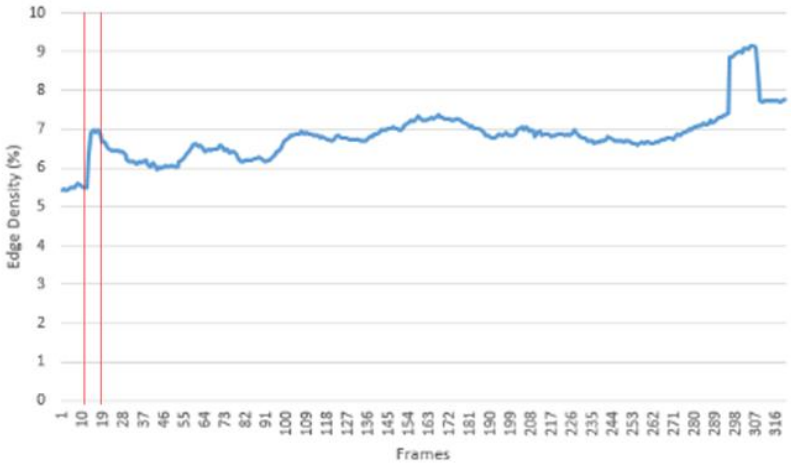
Since texts are composed of line segments and text regions contain rich edge information, an edge-based method is used to extract the text in video sequences. This dissertation shows firstly that the Canny edge detector is used to extract edge points.

The Canny edge detector is based on the specification of detection and localization criteria in a mathematical form. It is necessary to augment the original two criteria with multiple response measures in order to fully capture the intuition of good detection. A detector uses adaptive thresholding with hysteresis to eliminate streaking of edge contours. The thresholds are set according to the amount of noise in the image, as determined by a noise estimation scheme. This detector made use of several operator widths to cope with varying image signal-to-noise ratios, and operator outputs were combined using a method called operators were used to predicting the large operator responses. If the actual large operator outputs differ significantly from the predicted values, new edge points are marked. Therefore, it is possible to describe edges that occur at different scales, even if they are spatially coincident [59].

In the video sequences, all frames are acquired images and transform them into grayscale images. There are several ways to convert color images into grayscale images. This dissertation uses Y=0.299R+0.587G+0.11B to accomplish this. Where Y is the intensity value and R, G, B are the values on red, green, blue channels of the pixels. The Canny edge detector is applied to each grayscale image yielding an edge map. And then, the text edge density is computed in edge map images. The text edge density defines as the number of detected edge pixels in a frame over the total number of pixels in the frame [60].

Since the text presents many edges, the frame included the overlay text has significant changes in the text density than the frame not overlaid the text as shown in Fig. 3.2. The period between the vertical red lines, in other words from the frame 13 to frame 20 like Fig. 3.2(d) has sharply different from the previous frame and frame 21 has little different of the edge density of frame 22. As a result, the beginning frame becomes the frame 21.

(a) frame 1 (b) frame 13 (c) frame 25



(d) the plot of edge density in whole video sequences

Fig. 3.2. The result of edge density using the Canny edge detector

3.3.2 Method by the Harris Conner Detector

Corner points are the image features that are usually more salient and robust than the edge for pattern representation. A corner can be defined as the intersection of two edges or a point where there are two dominant and different edge directions in a local neighborhood of the point [6].

The corner points are the essential features by viewing the three-fold advantages for text detection. First, Corners are frequent and essential patterns in text regions. As an image feature, the corner is more stable and robust than other low-level features. Therefore, the impact of background noises can be eliminated to a large extent. Second, The distributions of corner points in text regions are usually more orderly in comparison to the nontext regions. Therefore, the unordered nontext corner points can be filtered out. At last, the usage of corner points generates more flexible and efficient criteria, under which the margin between the text and nontext regions in the feature space is discriminative [61].

In this dissertation, the second method is the Harris corner detector to extract the corner points. The Harris corner detector is a popular interest point detector, since its strong invariance to rotation, scale, illumination variation, and image noise The method is based on the local autocorrelation function of a signal, which measures the local change of the signal with patches shifted by a small amount in different directions [6].

After extracting the corner points, it is necessary to compute the shape properties of the regions containing corner points. As a result, the detector can make the decision to accept the regions as text or not. First, to do this, this dissertation performs image morphology dilation on the binary corner images. The separate corner points that are close to each other can be merged into a whole region. In the text, the presence of corner points is dense because characters do not appear alone but together with other characters and usually regularly placed in a horizontal string. Therefore, the text can be effectively detected by figuring out the shape properties of the detected regions [61].

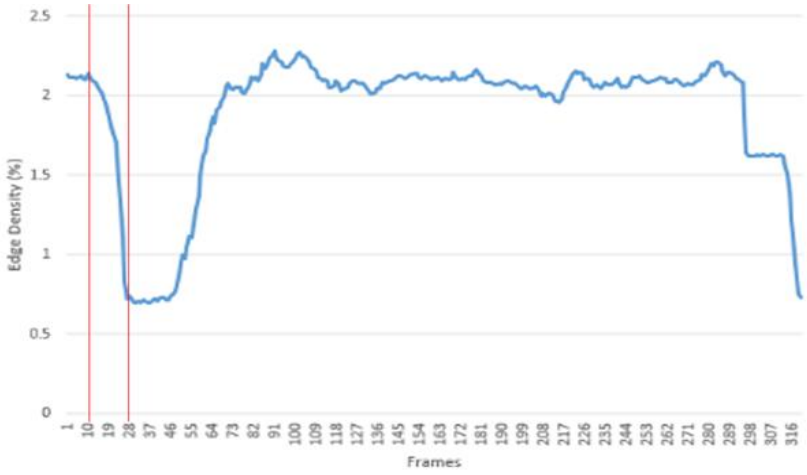
This dissertation uses the five regions properties, such as area, saturation, orientation, aspect ratio, and position, as the features to describe the text regions. First,

the area of a region is defined as the number of foreground pixels in the region enclosed by a rectangle bounding box. The small regions generated by the disorderly corner points can be easily filtered out according to the area measurement. Second, the saturation specifies the proportion of the foreground pixels in the bounding box that also belong to the region. This feature is very important for the cases where the nontext corner points can also generate regions with relatively large area values [61].

Third, Orientation is defined as the angle between the x-axis and major axis of the ellipse that has the same second moments as the regions, which range from −90 degree to 90 degree. Fourth, the aspect ratio of a bounding box is defined as the ratio of its width to its height. And last, the position information can be used to locate the text regions with a specific type and style. For example, the position is important to differentiate the overlay text from other text regions because overlay texts are usually superimposed by the production rules [61].

Fig. 3.3. shows the result images samples with detected corner points and edge density plot. The period between red vertical lines as shown in Fig. 3.3(d), in other words from the frame 12 to the frame 28, has sharply different from the edged density. Therefore, the beginning frame is frame 29.

(a) frame 1 (b) frame 13 (c) frame 25



(d) the plot of edge density in whole video sequences

Fig. 3.3. The result of edge density using the Harris corner detector

3.4 Video Clip Segmentation

By observing a large quantity of TV news programs, as shown in Fig. 3.4, the appearances and disappearances of the overlay text occur suddenly or slowly in most news videos. All frames in the video sequences do not contain in the overlay texts. By precisely locating the critical frame where each overlay text appears or disappears, the overlay text detection can be focused solely on the frame containing the overlay texts.

To do this end, this dissertation proposes the identification of the overlay text beginning frame. the method of the beginning frame identification is used the edge density based on the Canny edge detector. In the previous 3.2 and 3.3, this dissertation is compared to the performance of the beginning frame identification by the experiment between the Canny edge detector and the Harris corner detector. As a result, the appropriate processing method is proposed with the detection beginning frame with Canny edge.

Therefore, this dissertation considers the fact that the input video frames can be divided into three periods based on the beginning frame: non-text period, transition period, and text period as shown in Fig. 3.4. As a result, since the detection and recognition are limited only the text period, which consists of the frame superimposed onto the graphical text, the whole processing time is saved. And it helps to achieve a good result of the overlay text detection in news video sequences.

Fig. 3.5 shows the workflow of the video segmentation. In the video sequences, all frames are acquired images and transform them into grayscale images. The Canny edge detector is applied to each grayscale image yielding an edge map. And then, the text edge density is computed in edge map images. The text of edge density is compared between the current frame edge map image and the previous frame edge map image. If the variance of edge density is larger than the start threshold, the separation of the sub clip is started. If not, the next frame is compared. After starting, if the variance of edge density is larger than the end threshold, the separation is stopped. As a result, the input video sequences are separated the sub-clips, and the beginning frame is decided.

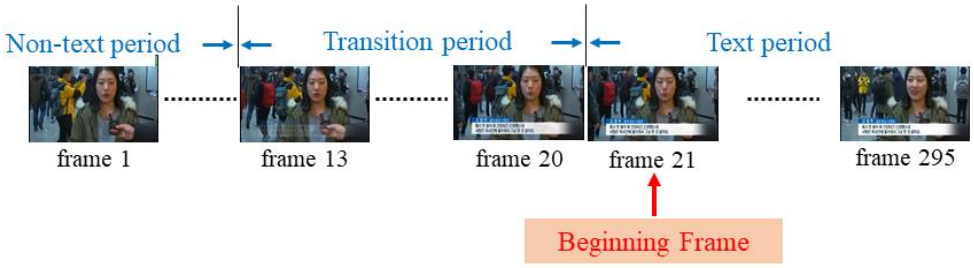


Fig. 3.4. The characteristics of overlay text in news video sequences

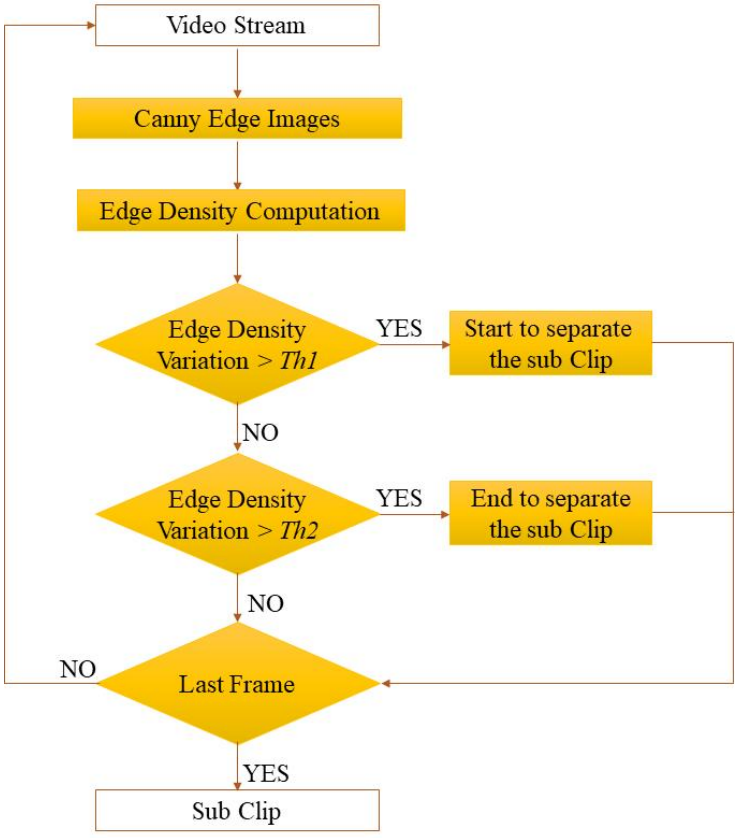


Fig. 3.5. The workflow for the video clip segmentation

Chapter 4

Name Text Line Localization

4.1 Introduction

To achieve OCR results accurately requires a good detection of the text regions in a video. The low resolution of the imagery, richness of the background and compression artifacts limit the detection accuracy that can be achieved in practice using existing text detection algorithms. Therefore, the text detection step must find a maximum amount of text, but also find exact coordinates of the boxes that contain text. Without an accurate surrounding box, the quality of text recognition is degraded, leading to poor performances.

Current video text detection approaches can be classified into two categories. One is detecting text regions individual frames independently. The other is utilizing the temporality of the video sequences. The former can be divided into the connected component-based methods, textures analysis based methods, and gradient edge-based methods. The latter is based on the fact that the overlay texts generally last at the same position for a few seconds [2, 4, 15, 18, 25, 62-64].

This dissertation uses the temporal analysis of the news videos to achieve a good accuracy of video text detection. And the method uses the rule-based characteristics in the production of the TV news program.

In the sub clip frames, the four representative frames (*R* frames) are selected and to detect the overlay text region, the logical AND operation executed on Canny edge maps of the four *R* frames as shown in Fig. 4.1. Since the same overlay text last in the same position for a few seconds or more, the representative frames selection reduces the processing time. The Multiple-Edge-Map image is acquired by the logical AND operation on the Canny edge map images of the selected representative frames. Based on these results, the overlay text region is detected by the number of black and white transition. The ROI(Region of Interest) text line mask which limits the region of interest in the whole text lines is obtained by the horizontal projection analysis of the detected overlay text region. At last, the overlay name text line is detected by applying the ROI text line mask image on one of the four *R* frames.

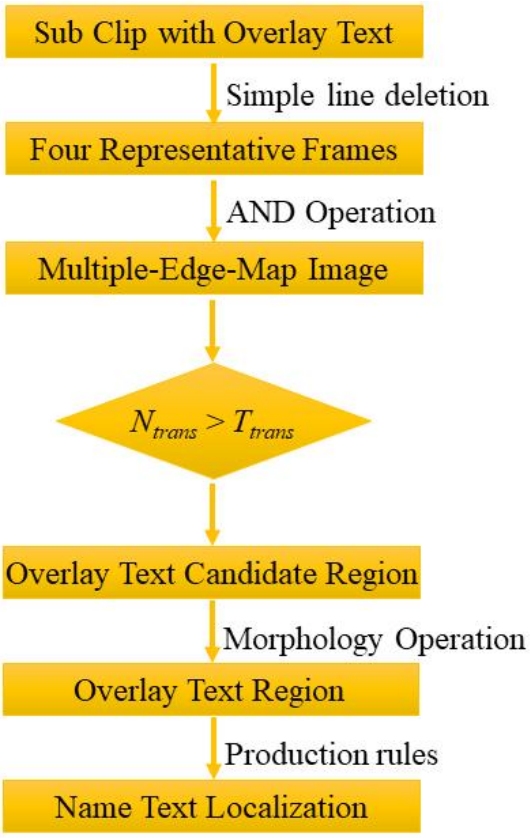


Fig. 4.1. The workflow of name text line localization

4.2 Representative Frames Selection

Since the broadcasting news videos are produced by professional, the overlay text of the TV news program uses the rule-based characteristics. In other words, by observing a large quantity of the TV news programs, the overlay text superimposed on most news videos has the following characteristics. The position of the overlay text is fixed; generally in the range of 1/2 from the bottom of the frame. The background of the text usually opaque or translucent matte, and in most case, the color of the background matte is ey-catching, such as white, blue, yellow, and so on. Colors of text character are often distinguishable from the background color. The overlay text is aligned horizontally. The size and font of the overlay texts in the same news video generally remain unchanged for a long term. And for readability in a complex scene, the same overlay text appears in the same position for a few seconds or more [6].

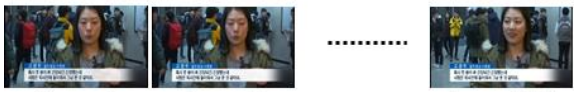
The representative frames (*R* frames) are selected in the sub-clips by the method proposed [60]. The text-frame choice is performed at association interval of two seconds for overlay text within the detected sub-clips. This can be an easy and economical resolution for video indexing applications that only needs keyword from video clips, instead of the whole text [19].

In general, a viewer needs 2 seconds or more to process a complex scene. Therefore, if videos are played f frames per second, the overlay text stays in a fixed location for at least $2f$ consecutive frames. Let k be the nearest integer that is not less than f. This dissertation defines every consecutive k frames to be one round. For example, the first round is made of frames from 1 to k, and the second round is made of frames from k+1 to $2k$, and so on. It can be shown that any $2k$ consecutive positive integers must 2 integers congruent to r modulo k that are k apart for any r=0, 1, ..., k-1. Therefore, any overlay text lasting for 2 seconds or more must appear frames $(m\text{-}1)k$ and mk for some positive integer m. As a result, the same text appears on the fixed position for every frame on the m^{th} round which is made of frames $(m\text{-}1)k$, $(m\text{-}1)k$+1, ..., mk [65].

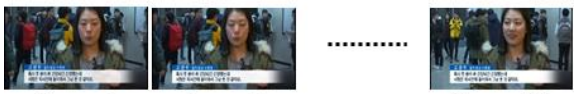
To simplify the calculation, about only 1st round, the four *R* frames are selected on frame 1, $\lfloor k/3 \rfloor$, $\lfloor 2k/3 \rfloor$, $\lfloor 3k/3 \rfloor$ as shown in Fig. 4.2. Because the same

overlay text is fixed in the same position for every consecutive k frames.

And then, the simple line deletion, horizontal and vertical, is used to remove long lines which are unlikely to be characters in the Canny edge result image of the four R frames. When the Canny edge image is scanned from the left to right and top to bottom, a horizontal line and vertical line is removed if its length exceeds the presumed width w and height h of a character. As a result, edge map images for four R frames are obtained as shown in Fig. 4.3.



(a) Whole sub clip frames



(b) 1st round frames

(c) Four representative frames (*R* frames)

Fig. 4.2. Representative frames (*R* frames) selection

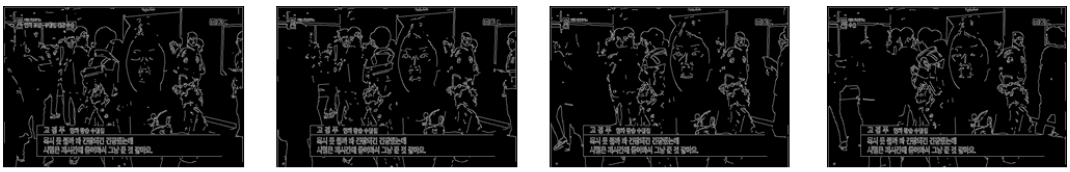


Fig. 4.3. The Four representative frames (*R* frames) edge images

4.3 Multiple-Edge-Map Image Generation

The logical AND operating on edge map images of four *R* frames is executed. The resulting image is called the Multiple-Edge-Map image. After AND operation, a position (*i*, *j*) becomes an edge pixel if all four edge images are edge at (*i*, *j*). By using AND, the overlay texts are kept if they are the same text on the same location for all four *R* frames Canny edge will be eliminated. Fig. 4.4 shows the Multiple-Edge-Map image. Therefore, most of the background edge pixels and the scrolling text on the bottom are removed, whereas the static overlay texts remain. Because the same overlay text appears in the same location for many successive frames, while the location of background edge pixels may differ in a few pixels.

The result well explains that the problem which the difficulty to distinguish whether the detected edges are really from overlay text is alleviated by multiple frame integration method.

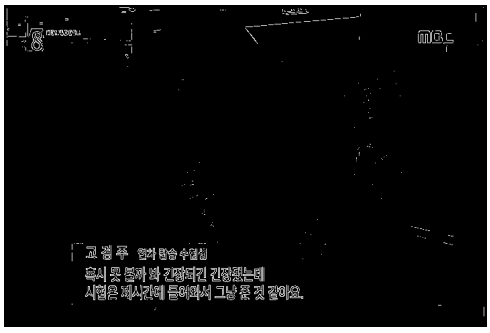


Fig. 4.4. The Multiple-Edge-Map image

4.4 Overlay Text Region Detection

To detection overlay text region, this dissertation implements three-stage steps. First, a rough text blob is obtained utilizing the number of black and white transitions column-wise and row-wise. Then, nontext noises and isolated noises are removed. At last, the morphological operation is applied for compensation to obtain the overlay text regions [65].

The overlay text candidate region is detected utilizing the number of the black and white transition. As shown in eq. (1), the value of N_{trans} can be obtained that a window of the presumed character size $w \times h$ slides from left to right and top and bottom on the Multiple-Edge-Amp image.

$$N_{trans} = \sum_{i=0}^{h-1} \left(\sum_{j=1}^{w-1} |b(i,j) - b(i,j-1)| \right) + \sum_{j=0}^{w-1} \left(\sum_{i=1}^{h-1} |b(i,j) - b(i-1,j)| \right) \quad (1)$$

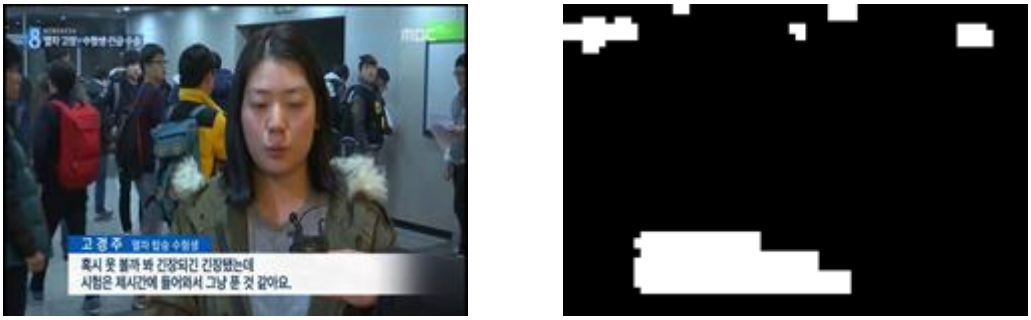
Where w and h are the width and height of the window and it is the presumed character size. And $b(\cdot)$ is a binary image. The value of N_{trans} represents the transitions from black to white or from white to black for every row and every column inside the window.

If N_{trans} is larger than threshold T_{trans}, this window masked. The union of all masked windows is the overlay text candidate region. The threshold T_{trans} depends on the character size and is obtained by $T_{trans}=\beta(w \times h)$ with β a constant which is empirically measured.

And then, every overlay text candidate region is examined from left to right and top to bottom for every masked pixel to remove the non-text ones. For a masked point located on (i, j) position, a horizontal line segment of length w comprising point on (i, j), ..., (i, j+w-1) will be eliminated if neither of these points is an edge point on the Multiple-Edge-Map. Simple connected component analysis is then followed to remove isolated pixels.

Due to various contrasts caused by different backgrounds in four R frames, the results of the Canny edge detector in four R frames of the same text in the same

location may differ in a few pixels. This causes characters to lose some pixels in the AND operation. To resolve the problem that characters lose some pixels in the AND operation, a morphological closing is applied first and then dilation is followed. A morphological closing with a horizontal structuring element of size $\lceil w/3 \rceil$ is used first to fill holes. And then, a dilation with structuring element of size $\lceil w/4 \rceil \times \lceil h/4 \rceil$ is used to connect the characters. The resulting image is the overlay text region as shown in Fig. 4.5 (b).



(a) 1st Representative frame (b) Overlay text region image

Fig. 4.5. The result of overlay text region detection

4.5 Name Text Line Detection and Localization

In general, many overlay texts can exist in one frame of the video. To detect a name text line, it is necessary not analyzing the whole overlay texts in the one frame. This dissertation constrains the detection region based on the news program production rules. TV content is produced by professionals. Many of the accepted production rules apply to TV contents.

As shown in Fig. 4.6, in the news interview video sequences, over a few lines, the story of interviewees is positioned at the bottom of the frame. And the interviewee's name and the title fix on the top line among the interviewee's first storylines and appear in the same position for a few seconds. Therefore, only the top of the first storylines is the region of interest (ROI).

To detect the ROI text line, at first, the horizontal projection histogram must be obtained by applying to the Multiple-Edge-Map image. To scan the result of the edge image from top to bottom, and count the number of the edge in a row can be gotten the horizontal projection histogram image as shown in eq. (2).

$$H_{hor}(i) = \sum_{j=0}^{w-1} b(i, \quad j), \qquad i = 0, 1, \cdots, h-1 \tag{2}$$

Fig. 4.6. The example of the broadcasting news interview video

Fig. 4.7 show the resulting image. This projection image is analyzed in the range of 1/2 from the bottom of the frame, making the process simpler. A horizontal line is discarded, when its projection value is lower than the minimum height. As a result, multi-lines can be segmented to single lines.

The first top area of the horizontal projection histogram in the half bottom region is selected as the region of interest. The start point and endpoint of ROI along the height (vertical) axis are applied to the overlay text region image. The result is the ROI text line mask image as shown in Fig. 4.8(a). At last, to apply to the one frame of four representative frame images yields the overlaid name text line image such as Fig. 4.8(b).

(a) Multiple-Edge-Map image (b) Horizontal projection image

Fig. 4.7. The result of horizontal projection detection

(a) ROI text line mask (b) Overlaid name text line image

Fig. 4.8. The result of overlaid name text line localization

Chapter 5

Personal Information Extraction

5.1 Introduction

The person identification based on the overlay text raises a lot of interest in the information research community. The identification using the Overlay Person Names (OPNs) has started to be investigated [2]. Since then the research area has raised a large amount of work, especially in face clustering tasks, face naming in captioned images, and recently, automatic naming within broadcast video [9-15]. These papers above deal with the application of the overlay text extraction and person identification. However, for content-based video search for the growing amount of video, it is necessary for the automatic video indexing and retrieval system.

To accomplish this goal, this dissertation discusses the person browsing system designed for the overlay text-based automatic person indexing database generation in the TV news videos as shown in Fig. 5.1. Especially, for broadcasted interview news video archives, the novelty of this dissertation presents how to build automatically indexing by identifying the named entities in the extracted overlay texts. Because the overlay texts

in the news interview video have characteristics to present name, age, occupation, address for the interviewee. Therefore, one of the contributions of this dissertation is that the overlay text-based name and title information in the interview video of the TV news program are valuable for building an information retrieval and data mining system.

The currently indexing database generation for retrieval is done manually by a person. An important human cost is induced by making the indexing database. And it is difficult to maintain the consistency of human subjective thoughts. Therefore, the need for automatic indexing database creation has raised. For this goal, this dissertation presents the method using the overlay text in video content. Since the overlay texts contained in the video represent rich and reliable information.

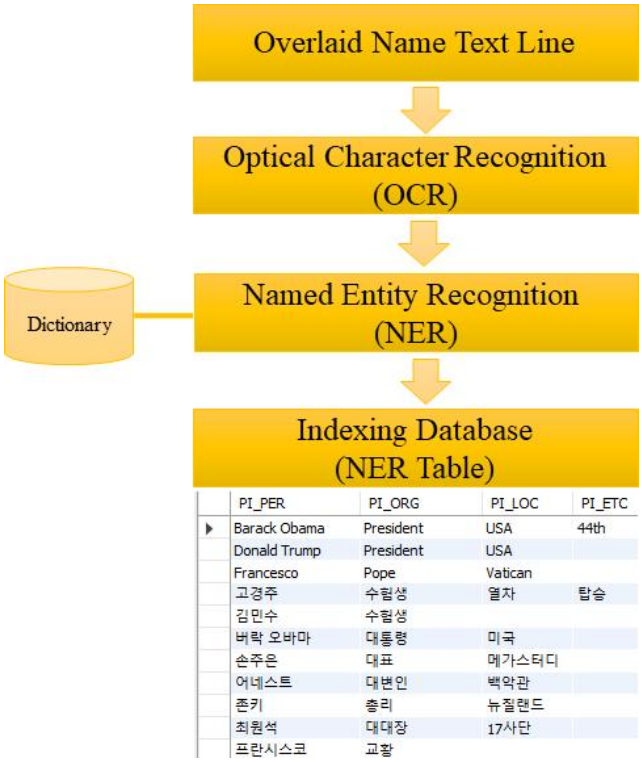


Fig. 5.1. The workflow of Personal information extraction

5.2 Text Recognition

The detected overlay name text using by the previously proposed method becomes the text string by the optical character recognition (OCR) processing. This dissertation do not intend to implement our own system. Instead, this dissertation can directly make use of existing state-of-art OCR software or packages such as Omnipage or the Scanwrox OCR software developed by Scandoft that provides a classification toolkit that can be easily integrated [54]. Therefore, this dissertation uses commercial software ABBYY FineReader [66] for OCR.

First, the binarized image is achieved with a gray threshold value derived from Otsu's method [67]. And then, the text string is obtained by the existing commercial software ABBYY FineReader as shown in Fig. 5.2.

5.3 Personal Information Extraction

The text string obtained by the OCR has important information such as person name, location, and organization of the interviewee. This information is helpful to organize the indexing and retrieval systems. Identifying and classifying the information not by manual but by automatic, it is very useful for broadcasted interview news video archives.

In most of the cases, to automatically extract information based on text use the automated natural language processing (NLP) tools. The natural language processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics. The goal is for computers to process or understand natural language in order to perform tasks like language translation and question answering. Information extraction methods try to identify portions of text that refer to a specific topic, by focusing on the appearance of instances of specific types of named entities such as person, organization, and location. In other words, people, places, and things play a crucial role in language, conveying the sentence's subject and often its object. Due to their importance, it's often useful when processing text to try to identify nouns and use them in applications. This task, often called either entity identification or named entity recognition (NER) is often

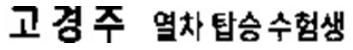
handled by a parser or chunk [57, 58, 68].

The term ‘named entity (NE)’ is generally considered to have originated at the Sixth Message Understanding Conference (MCU-6) held in 1995. Named entity recognition is a subtask of information extraction that seeks to locate and classify single words or multi-word expressions in text into pre-defined entity types such as the names of persons, organizations, and locations [69].

This dissertation look at how to perform the task of automatically identifying the specific type of named entities such as person, organization, and location. To perform the task, this dissertation uses stable and robust state-of-art the named entity recognition (NER) technology based on NLTK (Natural Language Tool Kit) package [70] and user-defined dictionaries.

The name text line mainly has a lot of content related to person name, occupation, and address of the interviewee. Therefore, in this dissertation, the implementation of NER recognizer can be used to recognize particularly the three classes: person name (tag: PER), organization (tag: ORG), and location (tag: LOC). The example of the result of extracting a person name and organization is shown in Fig. 5.3.

(a) Name text line image



(b) Name text line binary image

고경주 열차 탑승 수험생

(c) Text string obtained by ABBYY FineReader

Fig. 5.2. The example of a text recognition result

고경주 열차 탑승 수험생

(a) Text string obtained by OCR

고경주[PER]

수험생[ORG]

(b) NER result

Fig. 5.3. The example of a NER result of the name text line

5.4 Person Indexing Database

The person indexing database consists of the NER table, ClipInfo Table, Facethumb Table. Table 1 is an example of the NER table in the indexing database. By inserting this tagging result into each element of the NER table, the indexing database can be configured automatically.

And the face thumbnail of the Facethumb table is obtained in the first representative frames. We do not intend to implement our own system. Instead, this dissertation can directly make use of existing methods. Therefore, face are detected using OpenCV implementation to obtain the face thumbnail [71]. Therefore, the indexing database is automatically obtained by the information extraction from the text string and the face thumbnail.

Table 1. The example of the NER table in the indexing database

PER	ORG	LOC	ETC
프란치스코	교황		
어네스트	대변인	백악관	
버락 오바마	대통령	미국	

Chapter 6

Experimental Results and Analysis

6.1 Data Set

Since there was no standard data set for the proposed method, the experimented videos were captured in TV news program in Korea, and also were collected in Youtube clips. And the news interview video clip was edited manually in the captured and collected video sequences.

The video resolution was 720×480, and the frame rate was 29.97 frames per second. Each video clip may last more than 2 seconds, and the overlay text of these video clips included more than one overlay text in one frame. The presumed character size $w \times h$ was 15×15 pixels. The threshold of the N_{trans} was set to be 0.15 and the threshold of the horizontal projection histogram analysis was the presumed character width 15.

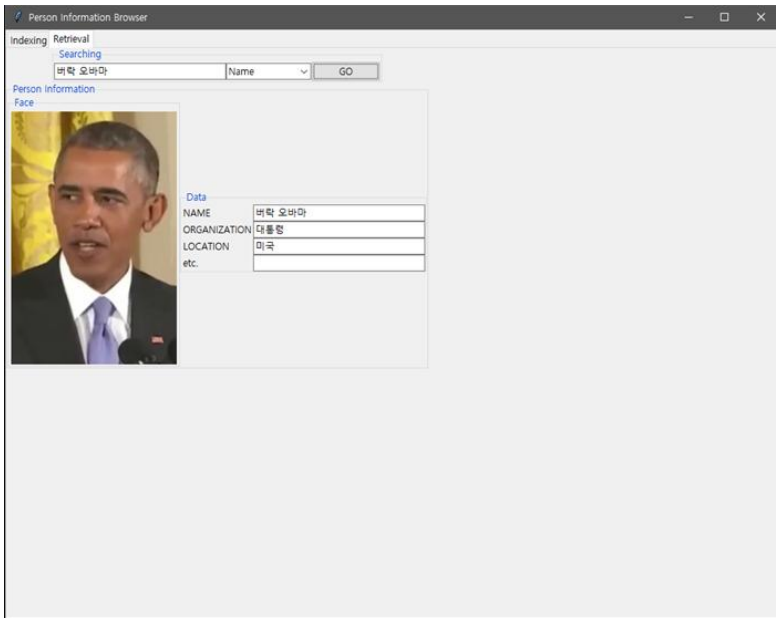
6.2 The Example of Implementation

The user interface of the person browser system using the proposed method is illustrated in Fig. 6.1. The system consists of two tabs: Indexing Tab and Retrieval Tab. This system implemented based on Python 3.6.6 and My-SQL 8.0 based on Windows 10 pro 64bit.

Fig. 6.1. (a) is the Indexing Tab user interface. The named entities in this Fig. were obtained by clicking the 'DETECTION' button at the desired video clip. And the indexing database is automatically generated by the 'SAVE' button at bottom of the indexing tab user interface.

Fig. 6.1. (b) is the Retrieval Tab user interface. At the 'Searching', if the user writes one of the named entities and clicks the 'GO' button, the result was shown at 'Data' about the personal information.

(a) The user interface for automatic personal information detection



(b) The user interface for retrieval

Fig. 6.1. Person browser system

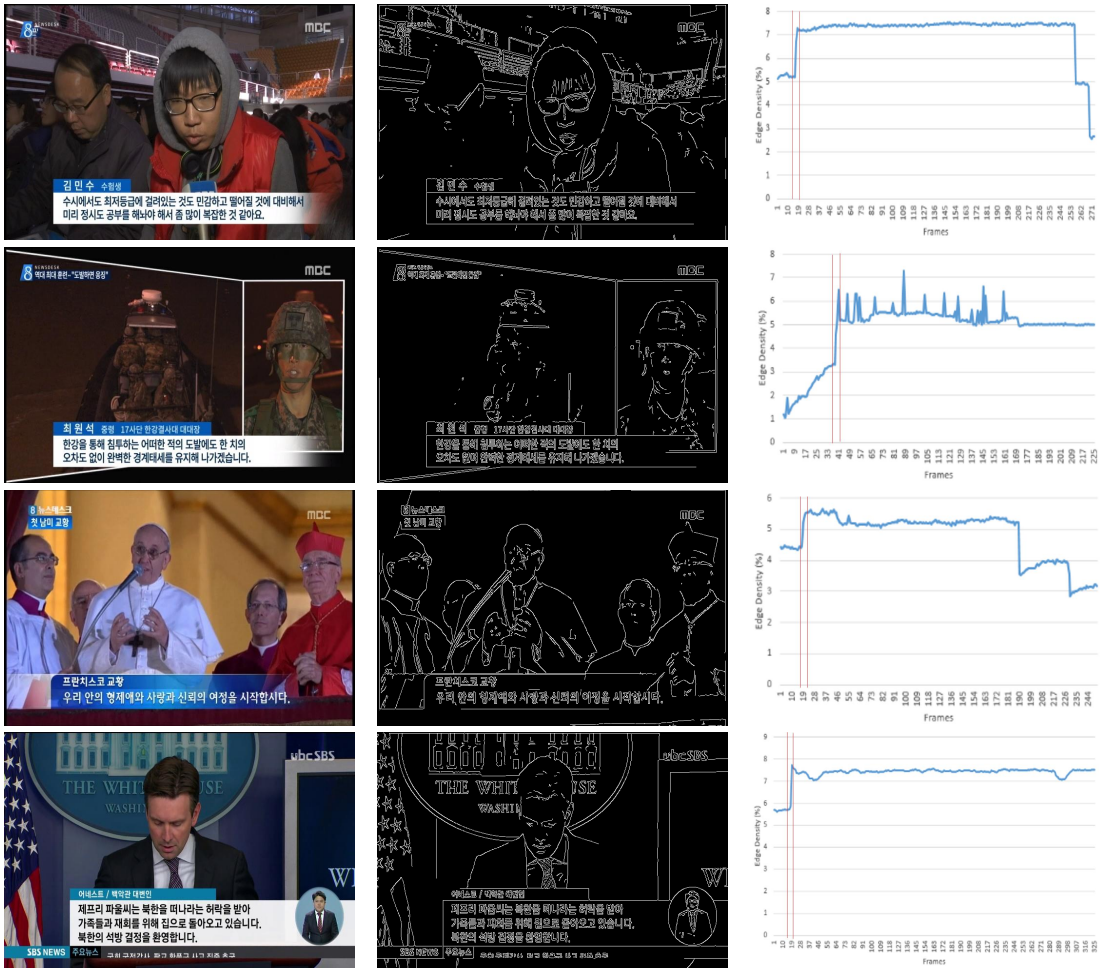
6.3 Beginning Frame Detection

6.3.1 The Beginning Frame Identification

Fig. 6.2 and Table 2 present the result of the overlay text beginning frame identification using the Canny edge detector. The two thresholds of the Canny edge detector, *i.e.*, T_{high} or T_{low} was decided based on empirical studies [60]. The reference value of abrupt difference among the frames was decided 0.03 by our empirical studies. By the beginning frame identification, the input videos were divided into three sub-periods; non-text period, transition period, and overlay text period.

Table 2. The results of the beginning frame identification

	Ground Truth		Canny edge Detector	
	Transition frame	Beginning frame	Transition frame	Beginning frame
TEST 1	16~24	25	16~21	22
TEST 2	38~47	48	35~43	44
TEST 3	18~25	26	18~27	28
TEST 4	18~24	25	16~22	23

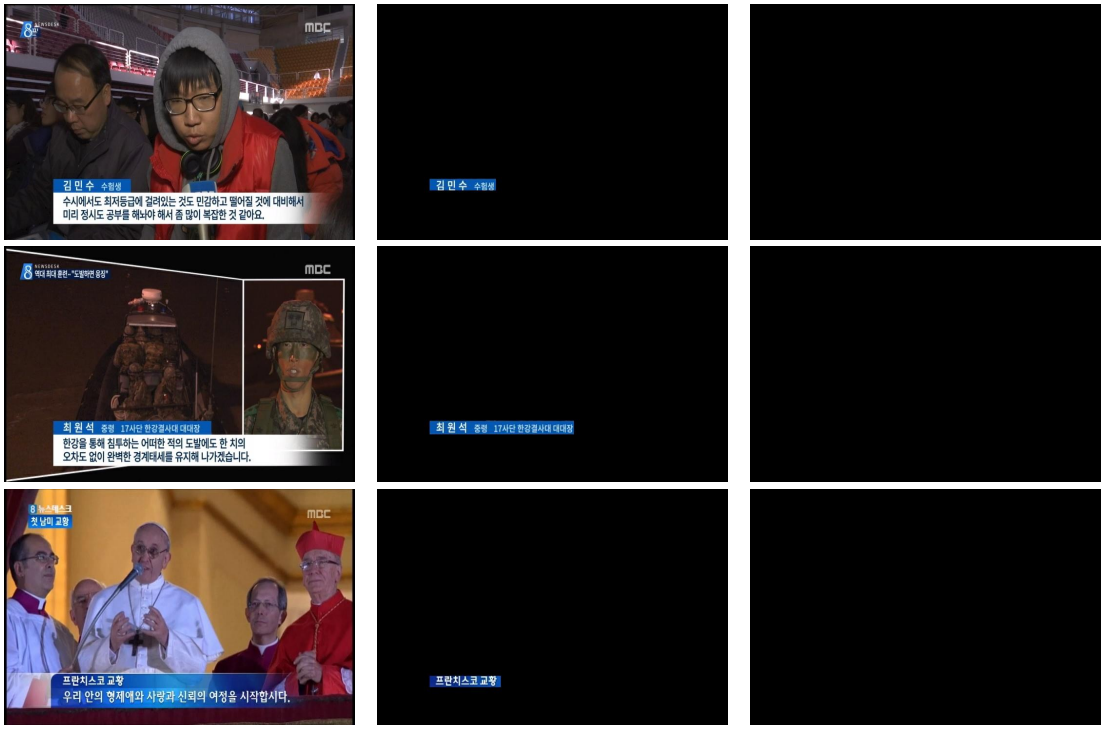


(a) Original image (b) Canny edge image (c) Plot of edge density

Fig. 6.2. The examples of the beginning frame identification

6.3.2 Usefulness of Beginning Frame Identification

To prove that the beginning frame identification is effective and useful, this dissertation was an experiment in two cased of using the beginning frame, and not. As shown in Fig. 6.3 (b), using the beginning frame properly detects the name text line. In contrast to, Fig. 6.3 (c) shows that not using the beginning frame identification method fail to detect the name text line. Therefore, for pre-processing of the overlay text detection, the usage of the beginning frame identification method helps to accurately detect the overlay text.



(a) Original image (b) Using beginning frame (c) Not using beginning frame

Fig. 6.3. The comparative experimental results of the usefulness of the beginning frame

6.3.3 Comparison of Text Beginning Frame Detection Methods

Fig. 6.4 and Table 3 showed the results of the overlay text beginning frame identification using the Canny edge detector and the Harris corner detector. Harris corner parameter k was 0.04.

In Fig. 6.4. the first column (a) is the original image superimposed onto the overlay text in video sequences. The second column (b) is the Canny edge image of (a), the third column (c) is the edge density plot using Canny edge detector in whole video sequences of (a). The fourth column (d) is the Harris corner image of (a), and the last column (e) is the edge density plot using Harris corner detector in whole video sequences of (a).

The result of the first row in Fig. 6.4 shows that the transition period is from frame 16 to frame 21 and the beginning frame is frame 22 using the Canny edge detector. Using the Harris corner detector, test 1 shows that the transition period is from frame 1 to frame 10 and the beginning frame is frame 11. The result of Canny edge is similar to that of the ground truth. On the contrary, the result of the Harris corner is wrong.

The result of the second row in Fig. 6.4 shows that the transition period is from frame 35 to frame 43 and the beginning frame is frame 44 using the Canny edge detector. Using the Harris corner detector, test 2 shows that the transition period is from frame 10 to frame 36 and the beginning frame is frame 17. The result of the Canny edge is similar to that of the ground truth. On the contrary, the result of the Harris corner is wrong.

The result of the third row in Fig. 6.4 shows that the transition period is from frame 18 to frame 27 and the beginning frame is the frame 28 using Canny edge detector. Using Harris corner detector, the test 3 shows that the transition period is from frame 29 to frame 47 and the beginning frame is the frame 48. The result of Canny edge is similar to that of the ground truth. On the contrary, the result of Harris corner is wrong.

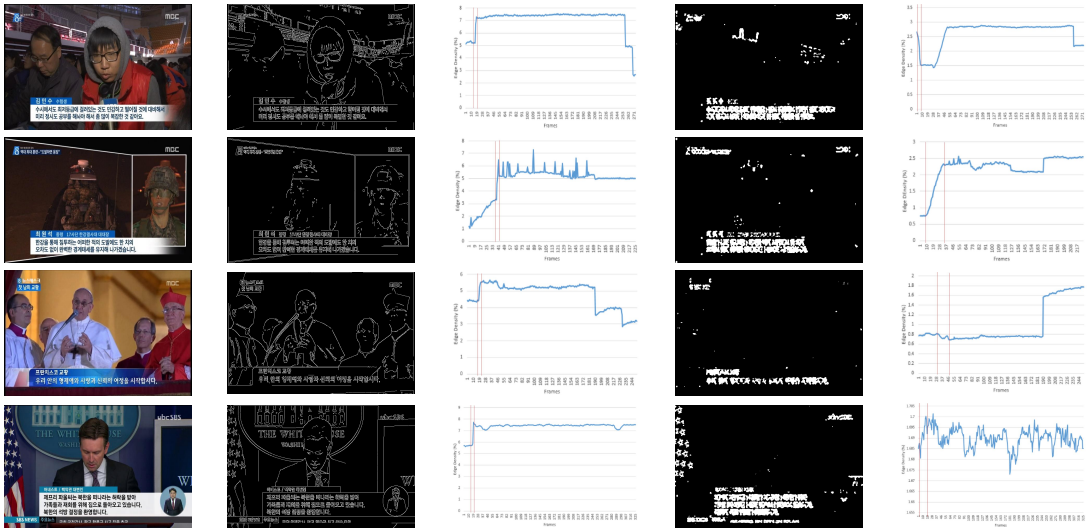
The result of the fourth row in Fig. 6.4 shows that the transition period is from frame 16 to frame 22 and the beginning frame is frame 23 using the Canny edge detector. Using the Harris corner detector, test 4 shows that the transition period is

from frame 2 to frame 19 and the beginning frame is frame 20. The result of the Canny edge is similar to that of the ground truth. On the contrary, the result of the Harris corner is wrong.

The Canny edge detector method relatively well detects the beginning frame than the Harris corner detector. Since the value of edge density using Harris corner is smaller than that using the Canny edge detector, the result shows that the Harris corner detector is more sensitivity than the Canny edge detector and its value deviation is bigger. As a result, the decision performance of the Harris corner detector is low and it cannot be used, whereas the beginning frame decision performance of the Canny edge detector is good.

Table 3. Comparison of the Canny edge detector with the Harris corner detector

	Ground Truth		Canny edge Detector		Harris corner detector	
	Transition frame	Beginning frame	Transition frame	Beginning frame	Transition frame	Beginning frame
TEST 1	16~24	25	16~21	22	1~10	11
TEST 2	38~47	48	35~43	44	10~36	37
TEST 3	18~25	26	18~27	28	29~47	48
TEST 4	18~24	25	16~22	23	2~19	20



(a) original image (b) Canny edge (c) Canny density (d) Harris corner (e) Harris density

Fig. 6.4. Comparison of beginning frame detection methods

6.4 Name Text Line Detection and Localization

Fig. 6.5 and Table 4 show the results of the overlaid name text line detection and localization. To evaluate the performance of the proposed algorithm, this dissertation shows the block level accuracy of the results of overlaid name text line detection and uses the precision and the recall as performance measures. *TP* (true positive) is the predicted positive block and *FN* (false negative) is the predicted negative block of the actual name text line. *FP* (false positive) is the predicted positive block and *TN* (true negative) is the predicted negative block of the actual non-name text line. The measure is calculated eq. (3) and eq. (4), and the result is shown in Table 4.

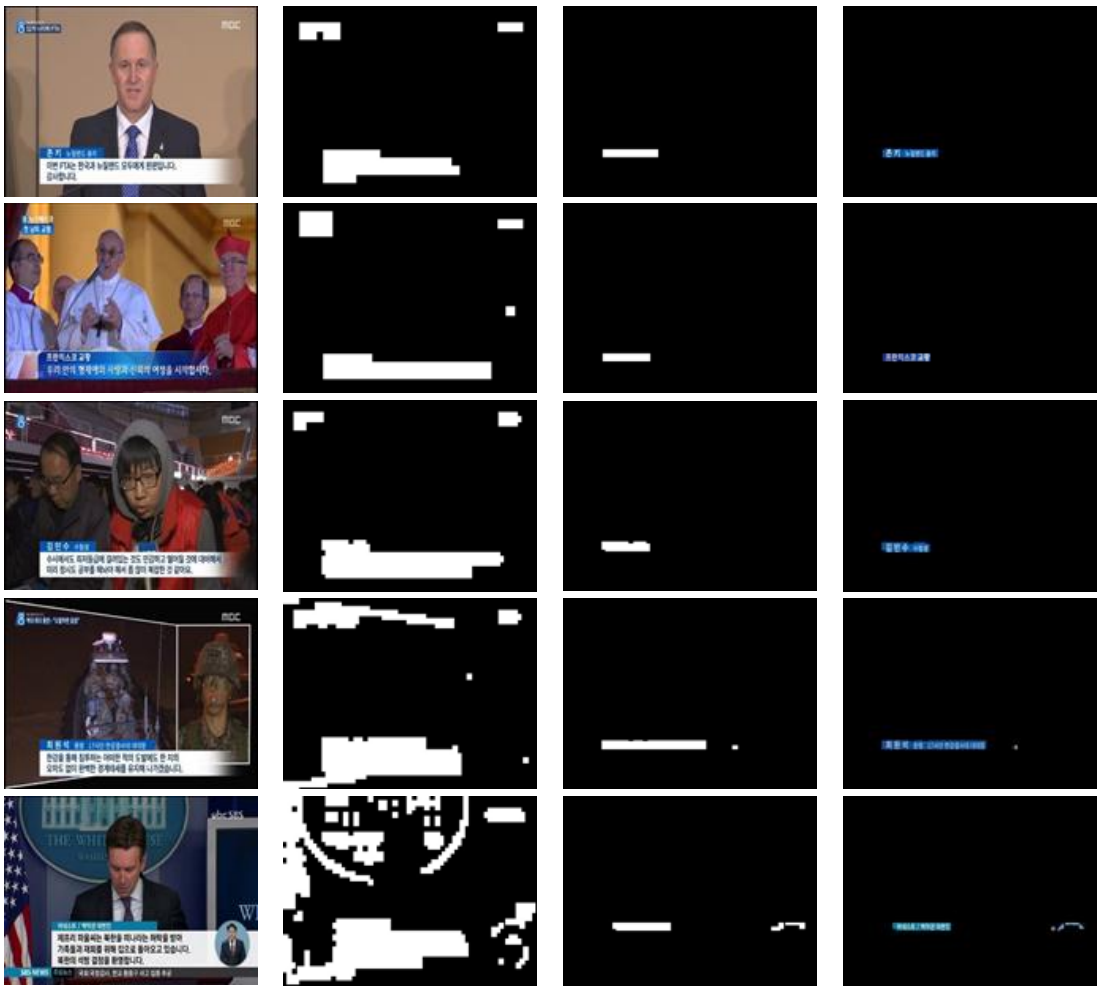
$$R_{pre} = \frac{TP}{TP + FP} \tag{3}$$

$$R_{rec} = \frac{TP}{TP + FN} \tag{4}$$

In Fig. 6.5, the (b) column is the result of the overlaid text region, the (c) column is the results of the ROI name line mask based on the horizontal projection analysis, the (d) column is the results of the detected overlay name text line. The proposed method accurately detect the position of the overlaid name text line in many examples. However, the fourth and the fifth-row results show that the false block in the overlay text name line was detected. In this case, many edges have in the nontext region and this area is detected as the text region. As a result, it is necessary to refine processes to remove this noise.

Table 4. The experimental results of name text line detection and localization

Method	Precision rate (%)	Recall rate (%)
Proposed method	66.67	100



(a) original image (b) Overlay text region (c) Name line mask (d) Name text line

Fig. 6.5. The result of name text line detection and localization

6.5 Name Text Recognition

First, the binarized image was achieved with a gray threshold valued derived from Otsu's method [67]. And then the text string was obtained by the commercial software ABBYY FineReader. Some example results shown in Fig. 6.6.

(a) Name text line (b) Binary image (c) Text string

Fig. 6.6. The examples of an experimental result of text recognition

6.6 Personal Information Extraction

For the named entity recognition (NER), this dissertation used the NLTK package. The named entities in this dissertation were defined as the person name (tag: PER), the organization (tag: ORG), and the location (tag: LOC). The NER results of text string obtained by the OCR showed in Table 5.

Table 5. The experimental results of named entity recognition

OCR results	Named Entity Recognition (NER)
프란치스코 교황	프란치스코[PER] 교황[ORG]
김민수 수험생	김민수[PER] 수험생[ORG]
어네스트 백악관 대변인	어네스트[PER] 백악관 대변인[ORG]

Chapter 7

Conclusion

7.1 Conclusion

In general, the overlay name text line of the news interview video contains person name, age, job, and so on. The name and title information in the interview video of the TV news program are valuable for building an indexing and retrieval system. Therefore, this dissertation presents the framework designed for the automatic person indexing and retrieval of the broadcasted news interview video sequences.

The currently indexing database generation for retrieval is done manually by a person. An important human cost is induced by making the indexing database. And it is difficult to maintain the consistency of human subjective thoughts. Therefore, the need for automatic indexing database creating has raised. This dissertation presents the method using the overlay text in video content to overcome the two problems.

The proposed methodology proposes a novel approach to extract meaningful content information from video by collaborative integration of image understanding and natural language processing. And the proposed methods are based on many of the

accepted production rules of the TV news and the temporality of the video sequences.

The proposed framework consists of the text detection module, the face detection module, and the person indexing database module. For the preprocessing step, the proposed system makes the sub-clip based on the beginning frame for only focusing on the frames with overlay text. In the text detection module, the system executives overlay text detection and separates the name text line. And the system processes detection and extraction of the overlay text, and text recognition by optical character recognition (OCR). In the face detection module, the face thumbnail is extracted. The face detection module makes the representative thumbnail of the interviewee. And the person indexing module generates automatically the index metadata by named entity recognition (NER). And finally, a person indexing database is automatically made by combining the recognized text with the face thumbnail.

Therefore, the proposed system enables the automatic video indexing and retrieval as well as the content-based video search in video portal and digital archives. In addition, the successful results of personal information extraction reveal that the proposed methodology of integrated use of image understanding techniques and natural language processing technique is headed in the right direction to achieve our goal of accessing real contents of multimedia information.

7.2 Future Research Direction

This dissertation presents the framework designed for the automatic person indexing and retrieval of the broadcasted news interview video sequences. The successful results of personal information extraction reveal that the proposed methodology of integrated use of image understanding techniques and natural language processing technique is headed in the right direction. However, there are still many problems to be further studied and discussed. The future works will be studied in the following aspects.

This dissertation constraints in the news interview video and have too many assumptions are made in this work, such as the position of overlay text, duration, background, and so on. Therefore, it will be necessary to develop the method for

processing the general broadcasting video genres. And many assumptions should be solved by machine learning methodology.

In this dissertation, the named entities are defined as the person name, the organization, and the location. It will be necessary to expand the number of named entities. And to improve the performance of named entity recognition, the study based on the convolution neural networks will be needed.

This dissertation presents the association of personal information and face in the broadcasted news interview video sequences. To enhance the accuracy of the retrieval is needed more information. For example, though this dissertation uses only the overlay name text line, the other overlay texts can be used as the information of the interviewee. And the internet person information, such as Wikipedia or Naver person database, can be used to make the indexing table.

The clue of information uses only text information in this dissertation. The enhancement of the retrieval is needed to exploit the feature of the image. Therefore, this will help to find the exact video sequences in the archives.

This dissertation experiment on Korean broadcasting news video sequences. It will be necessary to develop for automatic language detection and information extraction in many different language videos.

PUBLICATIONS

A. Patent

[1] 뉴스 인터뷰 영상의 오버레이 텍스트 기반 인물 인덱싱 방법 및 장치 (Method and apparatus for person indexing based on the overlay text of the news interview video), 제 10-1911613호, 2018. 10. 18.

B. Journals

[1] Sanghee Lee and Kanghyun Jo, "Person browser system based on named entity recognition for the broadcasted news interview," *Journal of Control, Automation and Systems*, submitted, 2019.

[2] Sanghee Lee, Hansung park, Jungil Ahn, Yeonsang On, and Kanghyun Jo, "Overlay text graphic region extraction for video quality enhancement application," *Journal of Broadcast Engineering*, Vol. 18, No. 4, pp.559-571, 2013.

[3] Sanghee Lee, Jungil Ahn, and Kanghyun Jo, "Automatic name line detection for person indexing based on overlay text," *Journal of Multimedia and Information System*, Vol. 2, No. 1, pp.163-170, 2015.

[4] Sanghee Lee, Jingil Ahn, Youlkyeong Lee, and Kanghyun Jo, "Beginning frame and edge based name text localization in news interview videos," *Lecture Notes in Artificial Intelligence* (LNAI), Vol. 9773, pp.1-12, 2016.

[5] Sanghee Lee, Junil Ah, and Kanghyun Jo, "Comparison of text beginning frame detection methods in news video sequences," *Journal of Broadcast Engineering*, Vol. 21, No. 3, pp.307-318, 2016.

[6] Sanghee Lee, and Kanghyun Jo, "Entity detection for information retrieval in video stream," *Lecture Notes in Artificial Intelligence* (LNAI), Vol. 10956, pp.618-627, 2018.

C. Conferences

[1] Sanghee Lee, Hansung park, Jungil Ahn, and Kanghyun Jo, "Overlay text box localization for 2D-to-3D video conversion," *International Conference of Image Processing and Image Understanding*, pp. 1-5, 2013.

[2] Sanghee Lee, Jungil Ahn, and Kanghyun Jo, "Comparison of text beginning frame detection methods for robust overlay text recognition," *International Workshop on Advanced Image Technology*, 2016.

[3] Sanghee Lee, Jingil Ahn, Youlkyeong Lee, and Kanghyun Jo, "Beginning frame and edge based name text localization in news interview videos," *International Conference of Intelligent Computing*, pp.583-594, 2016.

[4] Sanghee Lee, and Kanghyun Jo, "Strategy for automatic person indexing and retrieval system in news interview video sequences," *International Conference of Human System Interaction*, pp.212-215, 2017.

[5] Sanghee Lee, and Kanghyun Jo, "Automatic person information extraction using overlay text in television news interview," *International Conference of Industrial Informatics*, pp.583-588, 2017.

[6] Sanghee Lee, and Kanghyun Jo, "Entity detection for information retrieval in video stream," *International Conference of Intelligent Computing*, pp.618-627, 2018.

BIBLIOGRAPHY

[1] Zohra Saidance, Chrostophe Garcia, "An automatic method video character segmentation," *International Conference Image Analysis and Recognition*, pp. 557-566, 2008.

[2] Toshio Sato, Takeo Kanade, Ellen K. Huges, Michael A. Smith, and Shinici Sato, "Video OCR: Indexing digital news libraries by recognition of superimposed caption", *Multimedia Systems*, issue 5, vol. 7, pp.385-395, January 1999.

[3] Xian-Sheng Hua, Liu Wenyin, Hong-Jiang Zhang, "An Automatic Performance Evaluation Protocol for Video Text Detection Algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 498-507, April 2004.

[4] Qixiang Ye, David Doermann, "Text detection and recognition in imagery: A survey," *IEEE Transactions on Pattern and Analysis and Machine Intelligence*, 2013.

[5] Cheolkon Jung and Joongkyu Kim, "Player information extraction for semantic annotation in golf videos," *IEEE Transactions on Broadcasting*, Vol. 55, No. 1, pp. 79-83, March 2009.

[6] Sanghee Lee, Jungil Ahn, and Kanghyun Jo, "Comparison of text beginning frame detection methods in news video sequences," *Journal of Broadcasting Engineering*, Vol.21, No. 3, pp.307-318, May 2016.

[7] Zohra Saidane and Chirstophe Garcia, "Robust binarization for video text recognition," *International Conference on Document Analysis and Recognition*, Vol. 2 pp.874-879, 2007.

[8] Sanghee Lee, Jungil Ahn, Youlkyeoung Lee, and Kanghyun Jo, "Beginning frame and edge based name text localization in news interview videos," *ICIC201*6, pp. 583-594, 2016.

[9] Paul Gay, Gregor Dupuy, Carole Lailler, Jean-Marc Odobez, Sylvain Meignier, and Paul Deleglise, "Comparison of two methods for unsupervised person identification in TV shows," *12th international workshop on content based multimedia indexing*, 2014.

[10] Phi The Pham, Tinne Tuytelaars, and Marie-Francine Mones, "Naming people in news videos with label propagation," *Proc. of ICME*, 2010.

[11] B. Jou, H. Li, G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang, "Structured exploration of who, what, when, and where in heterogeneous multimedia news source," *Proc. of ACM Multimedia*, 2013.

[12] J. Poignant, L. Besacier, V. B. Le, S. Rosset, and G. Quenot, "Unsupervised speaker identification in TV broadcast based on written names," *Proc. of Interspeech*, 2013.

[13] J. Poignant, H. Bredin, V. B. Le, L. Besacier, C. Barras, and G. Quenot, "Unsupervised speaker identification using overlay texts in TV broadcast," *Proc. of Interspeech,* 2012.

[14] M. Bendris, B.Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, and J. Martinet, "Unsupervised face identification in TV content using audio-visual sources,", *Proc. of CBMI*, 2013.

[15] Johan Poignant, Laurent Besacier, George Quenot, and Frank Thollard, "From text detection in videos to person identification," *International Conference on Multimedia and Expo*, pp. 854-859, 2013.

[16] Shin'ichi Satoh, Yuichi Nakamura, Takeo Kande, "Name-It: Naming and detecting faces in news videos," *Proc. of IEEE Multimedia*, 1999.

[17] Christian Wolf and Jean-Michel Jolion, "Extraction and recognition of artificial text in multimedia documents," *Formal Pattern Analysis and Applications*, Vol. 6, Issue 4, pp.309-326, February 2004.

[18] Zhujun Wang, Lei Yang, Xiaoyu Wu, and Ying Zhang, "A survey on video caption extraction technology," *International Conference on Multimedia Information Networking and Security*, pp. 713-716, 2012.

[19] Avinash N bhute and B.B. Mesharam, "Text based Approach for indexing and retrieval of image and videos: A Review," *Advance in Vision Computing: an International Journal (AVC)*, vol. 1, no. 1, pp. 27-38, March 2014.

[20] H. K. Kim, "Efficient automatic text location method and content-based indexing and structuring of video database," *Journal of Visual Communication and Image Representation*, Vol. 7, No. 4, pp.336-344, 1996.

[21] M. A. Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization," *Technical Report CMU-CS-95-186, Carnegie Mellon University*, July 1995, 1995.

[22] U. Cargi, S. Antani, and R. Kasturi, "Indexing text events in digital video database," *Proc. of International Conference on Pattern Recognition*, Vol. 1, pp.1482-1483, 1998.

[23] Y. K. Lim, S. H. Choi, and S. W. Lee, "Text extraction in MPEG compressed video for content-based indexing," *International Conference on Pattern Recognition*, pp.409-412, 2000.

[24] Zhong Ji, Jian Wang, and Yu-Ting Su "Text detection in video frames using hybrid features," *International Conference on Machine Learning and Cybernetics*, pp.318-322, 2009.

[25] Jing Zhang and Rangachar Kasturi, "Extracton of text objects in video documents: Recent progress," *IAPR Workshop on Document Analysis Systems*, pp. 5-17, 2008.

[26] A. K. Jain and B. Yu, "Automatic text location in images and video frames," *Pattern Recognition*, Vol. 32, No. 12, pp.2055-2076, 1998.

[27] C. Garcia, X. Apostolidis, "Text detection and segmentation in complex color images," *IEEE International Conference of Acoustics, Speech and Signal Processing*, pp.2326-2330, 2000.

[28] D. Karatzas, A. Antonacopoulos, "Text extraction from web images based on a split-and-merge segmentation method using colour perception," *IEEE International Conference of Pattern Recognition*, pp.634-637, 2004.

[29] D. Chen, J. M. Odobez, H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognition*, Vol. 37, No. 3, pp.596-608, 2004.

[30] N. Nikolaou and N. Papamarkos, "Color reduction for complex document images," *International Journal of Imaging Systems and Technology*, Vol. 19, pp.14-26, 2009.

[31] V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: An automatic system to detect and recognize text in images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, pp.1224-1229, 1999.

[32] T. Q. Phan, P. Shivakumara, and C. L. Tan, "Text detection in natural scenes using gradient vector flow-guided symmetry," *IEEE International Conference on Document Analysis and Recognition*, pp.126-130, 2011.

[33] Kwang In Kim, Keechul Jung, and Jin Hyung Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *Pattern Analysis and Machine Intelligence*, Vol. 25, No. 12, pp.1631-1639, 2003.

[34] D. Chen, J. M. Odobez, and H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognition*, pp.595-608, 2004.

[35] X. Qian, G. Liu, H. Wang, and R. Su, "Text detection, localization, and tracking in compressed video," *Signal Processing: Image Communication*, Vol. 22, pp.752-768, 2007.

[36] Z. Ji, J. Wang, and Yu-Ting Su, "Text detection in video frames hybrid features," *ICMLC*, pp.318-322, 2009.

[37] X. Peng, H. Cao, R. Prasad, and P. Natarajan, "Text extraction from video using conditional random fields," *ICDAR*, pp.1029-1033, 2011.

[38] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A hybrid system for text detection in video frames," *Document Analysis Systems, DAS*, pp.286-292, 2008.

[39] R. Lienhart and F. Stuber, "Automatic text recognition in digital videos," *Proc. of SPIE*, pp.180-188, 1996.

[40] H. Li, D. Doerman, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Transactions on Image Processing*, Vol. 9, No. 1, pp.147-156, 2000.

[41] M. Sawaki, H. Murase, and N. Hagita, "Automatic acquisition of Context-based image templates for degraded character recognition in scene images," *International Conference on Pattern Recognition*, Vol. 4, pp.15-18, 2000

[42] S. Antani, U. Gargo, D. Crandall, T. Gandhi, and R. Kasturi, "Extraction of text in video," *Technical Report of Department of Computer Science and Engineering, Penn. State University, CSE-99-016*, August 30, 1999.
[43] U. Gargi, D.Crandall, S. Antani, T.Ganhi, R. Keemer, and R. Kasturi, "A system for automatic text detection in video," *International Conference on Document Analysis and Recognition*, pp.29-31, 1999.
[44] A. N. Bhute, B. B. Meshram, "Novel approaches for performing clustering and classification tasks using graphs similarity techniques: A review," *Proceeding of International Conference on Computing, Communication and Information Technology Applications*, pp. 113-119, 2010.
[45] Alan eustace, "A fall spring-clean," *Official Google Blog*, 2011.
[46] SY. Liu, ZL. Cao, "Single frame characters extraction in complex chromatic image," *Journal of Tianjin University of Technology*, Vol. 23, pp.58-61, 2007.
[47] L. N. Sun and L. Y. Fan, "Video text detection, location and extraction," *Electronic Technology*, Vol. 22, pp.75-79, 2009.
[48] Wonjung Kim, "A new approach for overlay text detection and extraction from complex video scene," *IEEE Transactions on Image Processing*, Vol. 18, pp.401-411, 2009.
[49] D. Aboutajdine, S. Elfkihi, and A. Jibab, "Features extraction for text detection and location," *ISVC*, pp.1-4, 2010.
[50] P. Shivakumara, S, Bhowmick, B. Su, C. L. Tan, and U. Pal, "A new gradient based character segmentation method for video text recognition," *ICDAR*, pp.126-130, 2011.
[51] Q. Wang, LQ. Chen, and X. Liang, "Text extraction in video," *Computer Engineering and Applications*, Vol. 48, pp.177-178, 2012.
[52] P. Shivakumara, RP Sreedhar, TQ Phan, and S Lu, "Multi-oriented video scene text detection through Bayesian classification and boundary growing," *IEEE Transactions on Circuits and System for Video Technology*, 2012.
[53] Rainer Lienhart and Wolfgang Effelsberg, "Automatic text segmentation and text recognition for video indexing," *Multimedia Systems*, vol. 8, no. 1, pp. 69-81, 2000.
[54] Huiping Li and David Doermann, "Video indexing and retrieval based on recognized text," pp.245-248, 2002.
[55] Jawahar, C. V., Balakrishna Chennupati, Balamanohar Paluri, and Natarai Jammalamadaka, "Video retrieval based on textual queries," *Proceedings of the Thirteenth International Conference on Advanced Computing and Communication*s, Coimbastore, 2005.
[56] Cees G. M. Snoek, Bouke Huurnink, Laura Hollink, Maarten De Rijke, Guus Schreiber, and Marcel Worring, "Adding semantics to detector for video retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp.975-986, 2007.

[57] Pavlina Fragkou, “Text segmentation using named entity recognition and co-reference resolution in english and Greek texts,” *Cornell University*, *Information Retrieval,* arXiv:1610.09226, 2016.
[58] Grant S. Ingersoll, Thomas S. Morton, and Andrew L. Farris, “Taming Text, How to find, organize, and manipulate it,” *Manning*, December 2012.
[59] John Canny, “A computational approach to edge detection,” *IEEE transaction on Pattern Analysis and Machine Intelligence*, VOL. PAMI-8, NO. 6, November 1986.
[60] Sanghee Lee, Junghil Ahn, and Kanghyun Jo, “Automatic name line detection for person indexing based on overlay text,” *Journal of Multimedia and Information Systems*, Vol 2. No 1. pp. 163-170, March 2015.
[61] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Yuxiao Hu, Yuncai Liu, and Thomas S. Huang, “Text from corners: A novel approach to detect text and caption in videos,” *IEEE transaction on image processing*, VOL. 20, NO. 3, March 2011.
[62] Jiamin Xu, Palaiahnakote Shivakumara, Ton Lu, Trung Quy Phan, and Chew Lim Tan, “Graphics and scene text classification in video,” *International Conference on Pattern Recognition*, pp. 4714-4719, 2014.
[63] Hrishikesh B.Aradhye and Gregory K. Myers, “Exploiting video text “Events” for improved video text detection,” *ICDAR*, 2007.
[64] Chien-Chen Lee, Yu-Chun Chiang, Hau-Ming Huang, and Chun-Li Tsai, “A fast caption localization and detection for news videos,” *ICICI*, 2007.
[65] Shwu-huey Yen, Hsiao-wei Chang, Chia-jen Wang, Chun-wei Wang, “Robust news video text detection based on edges and line-deletion,” *WSEAS Transactions on Signal Processing*, Vol. 6, Issue 4, pp.186-195, October 2010.
[66] ABBYY cloud OCR SDK, www.orsdk.com
[67] Otus., N., “A thresholding selection method from gray level histogram”, *IEEE Transactions on System, Man, and Cybernetics*, 1979.
[68] Rohit sharma, “A complete tutorial for named entity recognition and extraction in natural language processing using neural nets,” Dec 6, 2018.
[69] Anne-Stine Ruud Husevåg, “Named entities in indexing: A case study of TV subtitles and metadata records,” *Networked Knowledge Organization Systems Workshop (NKOS 2016*, September 2016.
[70] NLTK Language toolkit, www.nltk.org
[71] OpenCV, www.opencv.org