TEXT AND MOVING OBJECTS SEGMENTATION IN VIDEO FILES

By

Ali Hussein Ahmed Alabed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

TEXT AND MOVING OBJECTS SEGMENTATION IN VIDEO FILES

By
Ali Hussein Ahmed Alabed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

Under the Supervision of

Dr. Omar Ahmed Nasr

…………………………………..

Associate Professor of Electronics and
Communications Engineering Department,
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

TEXT AND MOVING OBJECTS SEGMENTATION IN VIDEO FILES

By
Ali Hussein Ahmed Alabed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

Approved by the
Examining Committee

Dr. **Omar Ahmed Nasr**	Thesis Main Advisor
Prof. Dr. **Mohsen Rashwan**	Internal Examiner
Prof. Dr. **Sherif Abdou** Faculty of Computers and Information, Cairo University	External Examiner

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2018

Engineer's Name:	Ali Hussein Ahmed Alabed	
Date of Birth:	1/1/1987	
Nationality:	Yemeni	
E-mail:	ali.h.alabed@gmail.com	
Phone:	01158807466	
Address:	Faisal. Giza	
Registration Date:	1/10/2014	
Awarding Date:	2018	
Degree:	Master of Science	
Department:	Electronics and Communications Engineering	

Supervisors:

Dr. Omar Ahmed Nasr

Examiners:

Prof. Sherif Abdou (External examiner)
Faculty of Computers and Information,
Cairo University
Prof. Mohsen Rashwan (Internal examiner)
Dr. Omar Ahmed Nasr (Thesis main advisor)

Title of Thesis:

TEXT AND MOVING OBJECTS SEGMENTATION IN VIDEO FILES

Key Words:
Segmentation, downsampling, reconstruction, caption text, text matching.

Summary:

In this thesis, we focus on addressing three problems related to information extraction in the video. The first problem is fast segmentation for the “dominant” object in the video file, the second is the extraction of text caption from news bars, and the third is finding key frames related to advertisements in videos. In the first part, we introduced the use of downsampling video frames to reduce the computation time of video object segmentation while maintaining a very high segmentation accuracy. In the second part, we proposed a system to extract text caption in the news bars for different Arabic TV channels. Using optical flow and Hough transforms, the system was able to detect and classify text region into three categories: horizontal scrolling, vertical scrolling and static region. For horizontal scrolling text, the system constructed complete sentences even if the sentence spans more than one frame. In the last part, we introduce a very fast search algorithm for advertisements in the videos. The algorithm first indexes the video using some video features in a KD-tree, then a binary search is performed on the tree. All algorithms were tested, and the accuracy of the algorithms showed their suitability for practical applications.

Acknowledgments

First and foremost, all thanks and praise are due to ALLAH Almighty, the absolute source of knowledge and wisdom, without His guidance and help nothing of this work would have been achieved.

I would like to express my deepest gratitude to Dr. Omar A. Nasr, Assistant Professor of electronics and communication engineering, Faculty of Engineering, Cairo University, for giving me the opportunity to pursue the Master of Science Degree under his guidance and support. In fact, he has enlightened for me the way of sound scientific research by giving me valuable information, directions, and encouragement, and has dedicated a lot of his precious time and effort to the whole work with the result that this thesis has seen the light.

I would like to express my sincere gratitude to Dr. Elsayed Hemayed, Professor of computer engineer; computer Engineering Dept., Faculty of Engineering, Cairo University for the continuous support he gave me during my research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I would like also to thank Dr.Mohsen Rashwan, Dr.Sherif Abdou and the rest of the research team working in the Arabic Media Mining project. They gave me a warm and encouraging research environment, and their comments during our meetings helped a lot in enhancing the work of this thesis.

Finally, the author would like to thank his parents, as well as his family, his sisters and his brothers for their continuous support and believing in him even through hardest times away from home. To them the author says thank you, but he knows that no words of appreciation could be sufficient. The author would not have achieved this work without their help.

Table of Content

List of Tables

List of Figures

List of Symbols and Abbreviations

AMVO	:	Average of Median Value of Optical Flow.
BF	:	Brute Force matcher.
BRIEF	:	Binary Robust Independent Elementary Features.
BRISK	:	Binary Robust Invariant Scalable Keypoints.
DAG	:	Directed Acyclic Graph.
FAST	:	Features from Accelerated Segment Test.
GMM	:	Gaussian mixture models
HW	:	Hard Ware.
KAZE	:	Japanese Word that Means Wind.
OCR	:	Optical Character Recognition.
ORB	:	Oriented Fast and Rotated Brief.
SIFT	:	Scale-Invariant Feature Transform.
SURF	:	Speeded-Up Robust Features.
VMRF	:	Video-Based Markov Random Field.

Abstract

Video files are full of rich content of different formats. The moving objects in the video file convey information to the user. The caption text in the video usually contains a summary about the video file content. For example, in the news, the caption text in the news bar usually contains information about the current piece of news being read by the broadcaster. The speech of the broadcaster also contains some information to be conveyed.

In this thesis, we focus on addressing three problems related to information extraction in the video. The first problem is fast segmentation for the “dominant” object in the video file, the second is the extraction of text caption from news bars, and the third is finding key frames related to advertisements in the videos.

In the first part of the thesis, we propose an enhancement of one of the most accurate techniques in the literature for dominant object video segmentation to reach almost the same segmentation accuracy with less number of computations. We propose downsampling the input video frames based on the speed of the objects in the video, and then a reconstruction of the segmentation result of the downsampled frames are done at the output video. The reconstruction is based on the optical flow information on the video frames. We show that at least 50% reduction in computation time is achieved with almost no loss in the segmentation results.

In the second part of the thesis, we propose a number of techniques to extract the text caption in the news bars for different Arabic TV channels. The techniques work for the horizontally scrolling text, the vertically scrolling text, and the static text in the news bars. The developed algorithms can extract the text even if multiple text region existed on video, and they can differentiate between different regions. In the case of horizontally scrolling text, we reconstruct the complete sentence that spans multiple video frames. The algorithms were tested for different Arabic TV channels with very different flavors of news bars. The results show that the algorithm can detect the text regions with an average accuracy of 97%.

In the last part of the thesis, we introduce a very fast search algorithm for advertisements in the videos. The algorithm first indexes the video using some video features in a KD-tree. Then a binary search is performed on the tree. The input image to the algorithm, which is the advertisement to look for in the video, should not be identical to the one in the video. We proposed using the ORB, and the AKAZE features in order to represent the videos. Searching for an image in a video of duration one hour took 0.227 seconds, which shows the effectiveness of the proposed indexing and search algorithms.

Chapter 1 : Introduction

The wide spread of the digital video contents through the Internet in one hand, and the availability of processing power through GPUs and large servers on the other hand, made it possible to apply computer vision algorithms in order to understand the video content. Hence, video analytics and understanding applications became viable in the last few years.

There are many problems related to computer vision. Researchers have addressed problems related to image recognition, face detection and recognition, activity recognition, and many other problems. There are different metrics that are measured to quantify the performance of each of these algorithms:

1- Accuracy: any computer vision algorithm should accurately do the required task. For example, in a face recognition algorithm, the accuracy of correctly detected faces should be high in order to practically use the system

2- Processing speed: many of the computer vision algorithms should be performed in real time. Hence, the speed of processing is a very important metric to consider

3- Complexity: the more complex the algorithm is, the higher the cost of the required HW to run the algorithm

In this theses, we address three computer vision problems, and we consider the previous three metrics in the developed algorithms. The problems are: segmentation of the moving objects, extraction of caption text, and searching of advertisements in video files. For moving object segmentation, we propose an enhancement to one of the most accurate techniques in the literature for dominant object video segmentation to achieve almost the same segmentation accuracy in less computation time. For caption extraction, we focus on extracting the layered text of three regions: horizontal scrolling news bars, vertical scrolling news snippets, and fixed (static) news message regions. For advertisements searching, we focus on searching on about the time where the advertisements appear on video.

1.1.Video Object Segmentation

Video object segmentation is used to get a pixel-level segmentation for the foreground object in a video. Object proposals can be foreground regions, are commonly used in this problem. There are two points of the object proposals. First, we need to generate a lot of object proposals for each video frame. The generation of object proposals take a long time, so we apply downsampling of video frames before the generation of object proposals. the computation time of generation of object proposals is reduced. Second, we need to keep the selected proposals consistent through all the selected frames. Spatiotemporal graph (DAG) is a perfect solution for this problem. The object proposals represented by nodes and temporal relationships between the proposals represented by edges. This problem can be solved by dynamic programming.

Video object segmentation is used to detect the primary moving object in the video then extract it from the background in all frames. Video object segmentation is a well-researched problem and is a prerequisite for a change of high-level vision applications, including content based video retrieval, activity understanding and targeted content replacement. Figure1.1, shows an illustration of this problem.

Several methods in the past, have been proposed to solve this problem, however, most of them are supervised (e.g. [5, 8, 10]), which the methods use manual segmentation in the first frame and prorogate them to the next frames. A most successful technique to solve this problem is to generate the object proposals for all video frames [2]. where the Object proposals are regions in frame which can be the foreground objects. There can be multiple object proposals for each video frame, therefore the most important problem is how to select the correct object proposals. The object proposals are ranked; the top proposals are usually not the object proposals we want. In this case, directed acyclic graph (DAG) method can be used to solve this problem [1].

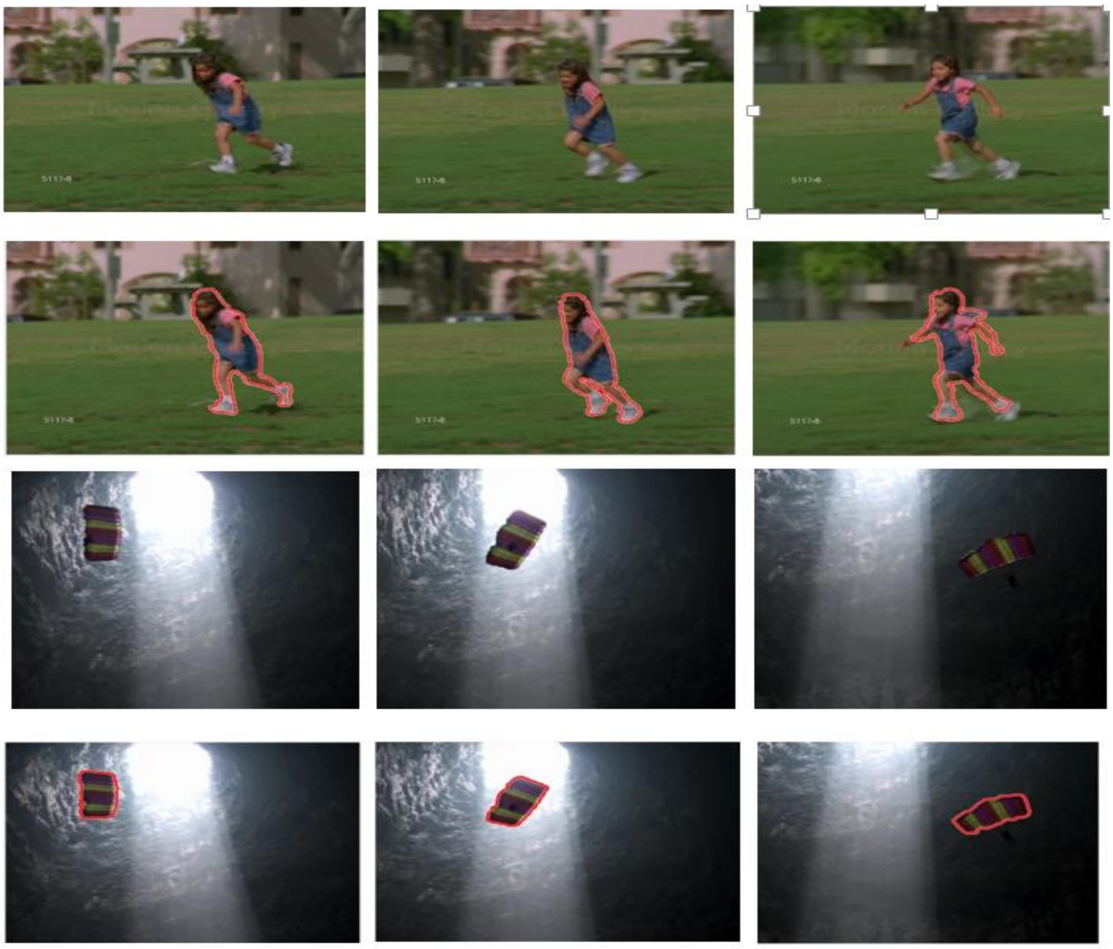

Figure 1.1: Video Object Segmentation illustration

In the moving object segmentation, we reduce complexity by downsampling video frames. Where the generation of object proposals for the selected frames takes small time compared to generation of object proposals of all original frames, then we apply DAG algorithm to select the correct primary object proposal, we propose reconstruction methods to get a primary moving object for unselected frames.

1.2.Extract a complete sentence from Arabic video news

Today, many TV channels upload its video content to online steaming platforms in order to target a larger audience. Automatic video understanding focuses on analyzing the visual, audio, and text content of the video in order to extract semantic information from the video. The text content in the video, either scene text, or caption text, is very useful in a video understanding task. text present in the video provides important information for indexing and retrieval, since text can provide a concise description of the stories and contextual clues of the objects presented in videos, particularly in news videos [3,4,6]. Text extraction from videos include five main phases: text detection, text localization, text follow-up, text segmentation, and text recognition.

Text detection to detect if the video frames contain text or not. Text localization gets rectangular bounding boxes of text lines. Text Follow-up gets accurate text boxes by applying empirical geometrical rules to verify the text boxes and eliminate false alarms. Text segmentation isolates text foreground pixels from its background. The text recognition converts the text foreground pixels into plain text [7].

Because of the importance of text in news videos, many research efforts have been given in text detection and localization. Progress has been made in this area, however the existing approach does not pay enough attention to the text appearance/disappearance patterns and also the different types of text. In this thesis, an edge-based method and Hough transform lines are proposed to detect, localize, and match text in Arabic news video to get a complete sentence.

In this thesis, we focus on extracting the information in the layered (caption) text in three different regions. The text in these regions has important market value because they usually carry the most important messages that the TV channel would like to convey to the audience. The text in these regions belongs to one of the following categories: (1) horizontal scrolling text, (2) vertical scrolling text, or (3) static text. Horizontal scrolling text carries a sequence of sentences that usually cannot be represented in one frame, and they are separated by a separator. The speed of scrolling, the length of the sentences, and the separator between sentences is different depending on TV channel. Horizontal scrolling text in news bars can appear in either news videos, or other TV programs. On the other hand, vertical scrolling and static layered texts usually appear in news videos. Vertical scrolling text is used when a short message is required to be conveyed to the audience for a short period of time. The same caption region is then used to show another message, and so on. Vertical scrolling text stays static for a period of time, it moves vertically, either up or down, and another message appears. Static text usually appears for a long period of time of a news program, and it is usually related directly to the current scene in the news video.