A SEMINAR REPORT

ON

"Tracking based Text detection and Recognition from web videos"

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY ,PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

THIRD YEAR OF ENGINEERING
(Computer Engineering)

BY

Miss.Deokar Harshada Tukaram
Exam Seat No: T150194217

UNDER THE GUIDANCE OF
Prof. Sawant V. V.

DEPARTMENT OF COMPUTER ENGINEERING
SVPM'S COLLEGE OF ENGINEERING
MALEGAON(Bk)
2018-19

SVPM'S COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING
SEMESTER II 2018-19

CERTIFICATE

This is to certify that the seminar work entitled

Tracking based Text Detection and Recognition from web videos

Submitted by

Miss.Deokar Harshada Tukaram

Exam Seat No: T150194217

It is a bonafide work carried out under the supervision of Prof. Sawant V. V. and it is submitted towards the partial fulfilment of the requirement of savitribai phule pune university,pune for the award of the degree of Third year of Engineering(Computer Engineering).

Prof.Sawant V. V.
(Guide)

Prof.Kumbhar H. R.
(HOD Computer Dept.)

Dr. S. M. Mukane
(Principal)

Place:
Date:

Abstract

Video is major source of visual or sensory data because of this intelligent analysis of video data is currently in wide demand. For multimedia understanding and retrieval video text extraction plays an important role. Most previous methods are conducted within individual frames. A few of recent research efforts, pay attention to text tracking using multiple frames. The generic Bayesian-based framework of tracking based Text Detection and Recognition (TTDAR) from web videos for embedded captions gives most probable . Which performs both tracking based text detection and tracking based text recognition in a single unified pipeline. This framework composed of three components, These are text tracking, tracking based text detection, and tracking based text recognition.

Acknowledgement

I am highly indebted to **Prof. Sawant V. V.** for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project.

I would like to express my special gratitude towards **Prof. Nimbalkar S. S.** for their kind co-operation and encouragement which help me in completion of this project.

I have taken efforts in this project. However, it would not have been possible without the kind support and help of **Prof. KUMBHAR H. R.** H.O.D Computer Department. I would like to extend my sincere thanks to all of them.

I would like to express my gratitude and thanks to Principal **Dr. MUKANE S. M.** to provide such a good environment for studies and other curricular activities and last but not least thank you for providing us such awesome LAB Facilities and services.

I am also thanks and appreciations to my colleagues who helped and co-operated with me in conducting the seminar by their active participation.

Deokar Harshada T.

Contents

List of Figures

List of Tables

Abbreviations

TTDAR	:	Tracking based Text Detection And Recognition
OCR	:	Optimal character Recognition
GCC	:	Gray-scale Consistency constraint
MAD	:	Mean absolute difference
SSD	:	Sum of Square difference
MSE	:	Mean Square error
SIFT	:	Scale Invariant Feature Transform
MAP	:	Maximum a posteriori

Chapter 1

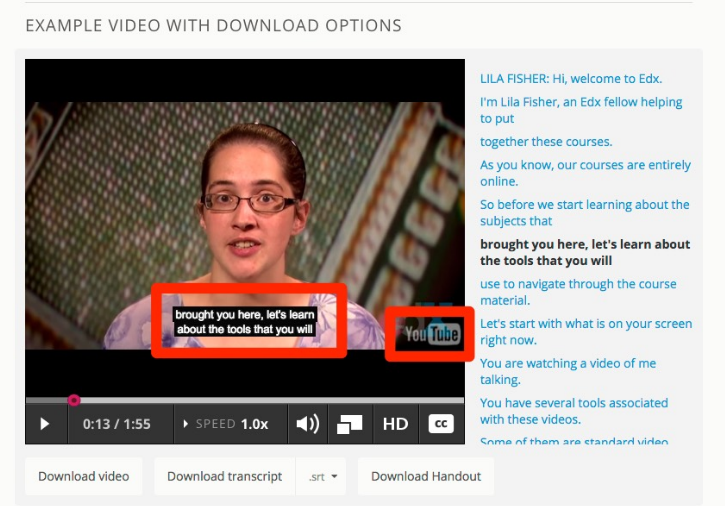
INTRODUCTION

1.1 Introduction

Text In Video :

There are two main types of text in the video. First is Caption text and second is Scene text [2]. Caption text is again classified into Layered caption text and Embedded caption text. Caption text gives a high -level overview and good directivity of the semantic information in subtitles, annotations and captions of video, while scene text is naturally embedded within objects. These are part of the camera- based images?

Embedded Caption Text In Web Videos :



EXAMPLE VIDEO WITH DOWNLOAD OPTIONS

brought you here, let's learn about the tools that you will

LILA FISHER: Hi, welcome to Edx.

I'm Lila Fisher, an Edx fellow helping to put

together these courses.

As you know, our courses are entirely online.

So before we start learning about the subjects that

brought you here, let's learn about the tools that you will

use to navigate through the course material.

Let's start with what is on your screen right now.

You are watching a video of me talking.

You have several tools associated with these videos.

Some of them are standard video

Figure 1.1: Embedded caption text in web video

Caption text is classified into layered caption text and embedded caption text, where the background of layered caption text is specifically designed background layer [2] and embedded caption text is placed on the frame. Embedded caption text and scene text are very difficult for detection, tracking, and recognition. Mainly embedded caption text appears in massive web videos with various challenges for detection and Recognition.

What Is Video Text Extraction :

It is the process of naturally fetching out and recognizing the content of the video. Automatic notification and distribution are very useful in real time by the contents of streaming video. video OCR, or video text recognition is a constructive tool to characterize the contents of video containing scene text (text that appears in the real scene of the video, such as text on street signs, nameplates, and billboards) And overlay text (text captions superimposed over the video imagery, such as in broadcast news programs).Digital Videos are widely used domestically and professionally [3] because of the easy availability of camcorders to mobile phones. People are increasingly making videos may be for personal use or commercial use, this is leading to the growing content of Video. While we can capture, compress, store, transmit and display video with great facility, It is an important activity that editing videos and manipulating them based on their content.

Text Tracking : Tracking is used to expand or contract the amount of text on a page by expanding or reducing the amount of space between letters. Tracking can be applied to an entire font or to a range of text. There are many methods for text tracking. Most of the methods are related to tracking with detection, in that detected text positions are used to track text across consecutive frames. these methods are again classified into text tracking with template matching, text tracking with particle filtering and text tracking with tracking-by-detection.

Tracking based Text Detection : Text detection methods mainly focus on detecting text in some frames or in each individual frame. It is the process of localizing the text in each video.Text tracking methods with multiple frames are consequently introduced in the detection process to reduce false alarms and to improve the accuracy of detection by exploiting temporal redundancy.

Tracking based Text Recognition : Text recognition is a segmenting text and recognizing it using optical character recognition method.Text recognition can be done by over-segmentation,merging,boundary refinement,and voting.over-segmentation and merging is the main issues in text recognition. to perform these two tasks using knowledge-based rules.A unied Bayesian-based framework is proposed for both tracking based text detection and tracking based text recognition from complex (web) videos for embedded captions for exploiting and sharing information between tracking and detection (tracking and recognition).

1.2 Motivation

Text objects embedded in video contain lots of correct knowledge related to the multimedia system. Also, they are a rich source of information for content-based indexing and retrieval applications.

Most of the common text in web videos includes some common properties. These are nothing but size, alignment, color, motion, way, edge, compression, [3] etc.why the text in videos becomes more popular or important day by day? And what is the need of extracting text from web videos?

- To understand an importance of text in videos.
- To understand the concept of text extraction.
- Text in videos give information about videos.

1.3 Objectives

The main objectives behind this topic are as follows:

- To identified types of text in web videos.
- Alert the user about a large amount of visual data present in a video on the and social network.
- Also user awareness about this text.
- Use various methods of tracking, tracking based text detection and recognition to extract text from web videos.
- Creating user awareness and providing video text extraction education.

1.4 Outcomes

The Outcomes regarding to this topic are as follows:

- We can extract text from web videos.
- We have understood some text detection and text recognition techniques specially in web videos.
- User awareness is generated.

1.5 Advantages

Some Advantages of this topic are as follows:

- It uses optimal character recognition (tesseract-OCR)tool.
- It is based on bayes theorem of probability hence,it gives most probable result.
- It gives video search in faster time.
- It works in single unified pipeline.

Chapter 2

LITERATURE REVIEW

2.1 Review of Literature

In general there are three types of text. They are as follows:

2.1.1 Layered Caption text:

Layered caption text is always printed on a specially designed background layer

2.1.2 Embedded Caption text: Embedded caption text embedded on the Frame.it is very difficult to extract text from embedded caption text.

2.1.3 Scene text: Scene text is nothing but part of camera-based images and it is naturally embedded trademarks, signboards and buildings etc. also it is very difficult to extract text from scene text.

2.2 Challenges

There are various challenges for detecting and recognizing embedded captions from web videos:

2.2.1 Complex Backgrounds

Variety of objects, scenes and graphics are dynamically changed because of these web videos have complex backgrounds. noises which are varied always follow Text detection results.

2.2.2 2.2.2 Varied colors:

The colors of text are dynamically changed because of video compression and background changes. Original color of superimposed text is a constant color, but actually displayed color of the caption text in consecutive frames has a wide Range.

2.2.3 Similar colors:

The colors of foreground text and backgrounds are similar in some cases. Text strokes are always interconnected with the background edges for embedded Captions .some text regions and characters are easily missed by text detector.

2.2.4 Low contrast:

Major part of text in web video frame is with low contrast and blur, hence web video frames have low resolution.

No	Paper Title	Concept	Advantages	Disadvantages
1	X.C Yin Text detection,tracking and recognition in video:A comprehensive survey [2]	Tracking with template matching	it involves feature extraction	for matching it uses MAD,SSD and MSE.hence computational burden is high.
2	X.C Yin Text detection,tracking and recognition in video:A comparepensive survey [2]	Tracking with particle filtering.	It recursively find state of system based on available observations.hence it is very useful in computer vision.	It is non linear filtering technique.
3	Shu Tian,Xu-cheng Yin, ”A Unified framework for tracking based text detection and recognition from web videos [1]	Tracking with tracking-by-detection	It uses appearance model for detection and motion model for tracking.	similarity calculation is very big issue.
4	Y.F pan A hybrid approach to detect and localize texts in natural scene images [3]	Sliding window based text detection method.	It gives good result in noisy images.	speed is slow because performance is sensitive to text alignment orientation.
5	Y.F pan A hybrid approach to detect and localize texts in natural scene images [3]	CC based Text detection method.	computational burden is low.	CC-based methods cannot segment text components accurately without prior [3] knowledge of text position and scale.
6	Y.F pan A hybrid approach to detect and localize texts in natural scene images [3]	hybrid method.	It gives good result because it is combination of region based and cc based method.	computational burden is high.

Table 2.1: Evolution of Various methods.

Chapter 3

SYSTEM OVERVIEW

3.1 System Architecture

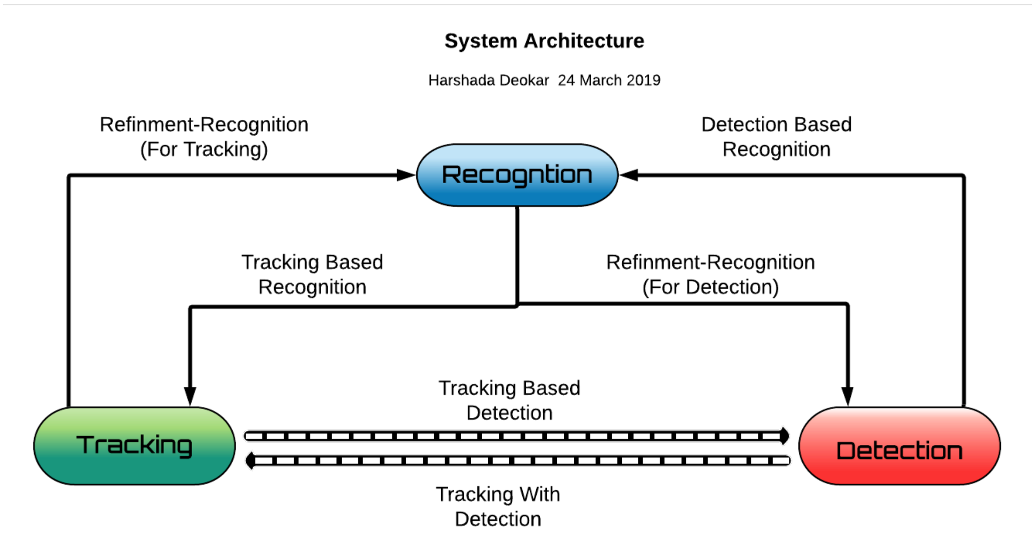

System Architecture
Harshada Deokar 24 March 2019
Refinment-Recognition (For Tracking)
Detection Based Recognition
Recogntion
Tracking Based Recognition
Refinment-Recognition (For Detection)
Tracking Based Detection
Tracking
Detection
Tracking With Detection

Figure 3.1: A Whole diagram for text detection,tracking and recognition in video

In general, the whole system of video text extraction includes several major components which are Detection, Recognition and Tracking. And their relations and interactions .

- Detection is the task of localizing the text in each video frame with bounding boxes. Tracking is the task of maintaining the integrity of the text location and tracking text across adjacent frames.
- Recognition involves segmenting text and recognizing it using Optical Character Recognition (OCR) techniques. Obviously, Recognition is performed on text regions detected from Detection results (Detection-based Recognition).

- Tracking uses the locations identied in the Detection step to track text (Tracking-with-Detection). In general, Detection is rst performed rst in each frame independently; then, the detection results in sequential frames can be integrated and enhanced based on the Tracking results (Tracking-based-Detection).
- Recognition can help verify the Tracking results (Renement-by-Recognition for Tracking), and also conrm the Detection results in some cases (Renement-by-Recognition for Detection). Meanwhile, Tracking can improve Recognition by fusing the recognition results over multiple frames (Tracking-based-Recognition).

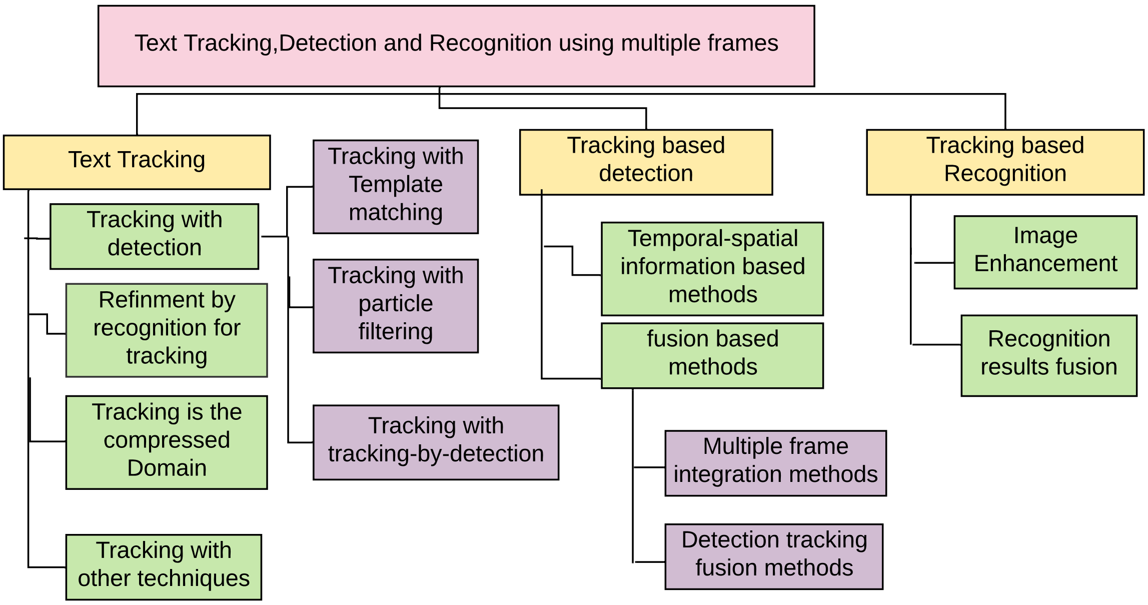

Text Tracking,Detection and Recognition using multiple frames
Text Tracking
Tracking with detection
Refinment by recognition for tracking
Tracking is the compressed Domain
Tracking with other techniques
Tracking with Template matching
Tracking with particle filtering
Tracking with tracking-by-detection
Tracking based detection
Temporal-spatial information based methods
fusion based methods
Multiple frame integration methods
Detection tracking fusion methods
Tracking based Recognition
Image Enhancement
Recognition results fusion

Figure 3.2: Different methods for Tracking, detection and Recognition

Chapter 4

BAYESIAN FORMULATION

given that detection and recognition result(prior probability) and the probability of observed data. it uses Maximum a posteriori (MAP) calculation.

All detections in frame t are

$$D_t = \{d_{t,1}, d_{t,2} \cdots \cdots d_{t,m_t}\}$$

where $d_{t,i}$ is the ith detection, m_t is the number of detections in frame t.

And all recognition results in frame t are

$$R_t = \{r_{t,1}, r_{t,2} \cdots \cdots r_{t,m_t}\}$$

where $r_{t,i}$ is the ith recognition in frame t.

All Trajectories generated from frame 1 to t as

$$T_t = \{T_{t,1}, T_{t,2} \cdots \cdots T_{t,n_t}\}$$

where n_t is the no of trajectories in frame t

$$T_{t,i} = \left\{d_{t,i,1}, d_{t,i,2} \cdots \cdots d_{t,i,p_{t,i}}\right\}$$

where $p_{t,i}$ is the length of trajectory $T_{t,i}$.

The target is X$_t^* = \arg\max P\left(X_t | X_t^0, T_{t-1}^*\right)$

According to total probability theorem,

$$P(X_t|X_t^0, T_{t-1}^*) = \sum_{T_t} P(X_t|T_t, X_t^0, T_{t-1}^*)P(T_t|X_t^0, T_{t-1}^*)$$

where X=D and X=R, D_t^0 is the output of text detector, R_t^0 is output of recognition. The most probable trajectories are obtained with

$$T_t^{*,0} = \arg\max P\left(T_t|X_t^0, T_{t-1}^*\right)$$

$$p(T_t^{*,0}|X_t^0, T_{t-1}^*) = 1,$$

$$p(T_t|X_t^0, T_{t-1}^*) = 0$$

T_t!=$T_t^{*,0}$

where $T_t^{*,0}$ initial value of trajectory

$$p(X_t|X_t^0, T_{t-1}^*) = p(X_t|T_t^{*,0}, X_t^0, T_{t-1}^*)$$

The final detection or recognition results in frame t are

X$_t^*$ $= \arg\max P\left(X_t|T_t^{*,0}, X_t^0, T_{t-1}^*\right)$

optimal trajectories in frame t T_t^* are updated as

$$T_t^* = \arg\max P\left(T_t|X_t^*, T_t^{*,0}\right)$$

these optimal trajectories will be used for optimization in frame t+1.

Chapter 5

Text Tracking

5.1 novel tracking-by-detection

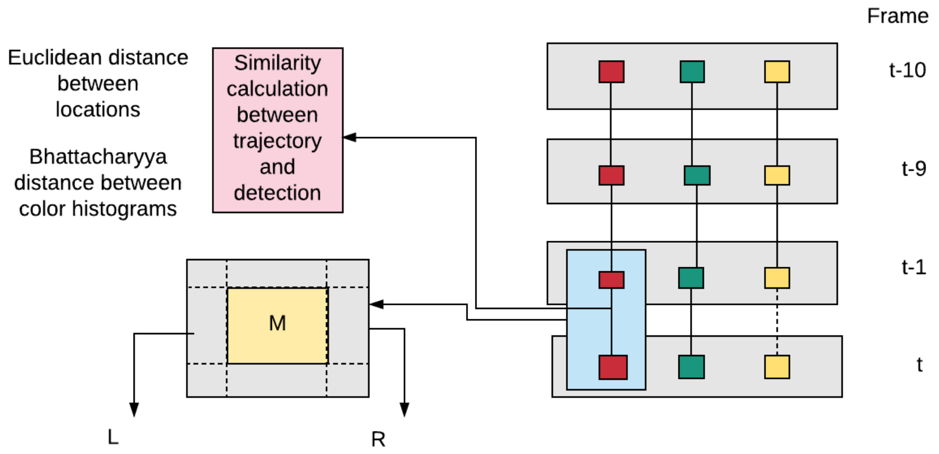

Euclidean distance between locations
Bhattacharyya distance between color histograms
Similarity calculation between trajectory and detection
Frame
t-10
t-9
t-1
t
M
L
R

Figure 5.1: Text tracking with a tracking-by-detection

In novel tracking by detection two models are introduced for text tracking,appearance model [6] for matching and the motion model for tracking are used to link the detections into trajectories.

5.2 Similarity Calculation

To compute the assignments between trajectories with detections hungarian algorithm is used.There are two types of similarities appearance similarity and location similarity.

Total similarity is combination of appearance and location similarity where s_a=appearance similarity s_l=location similarityto

$$S\left(T^*_{t-1,i},d^0_{t,j}\right)=S_a(T^*_{t-1,i},d^0_{t,j}).S_l(T^*_{t-1,i},d^0_{t,j})$$

Appearance Smilarity is

$$S_a(T^*_{t-1,i},d^0_{t,j})=exp\frac{(DB(H(d^*_{t-1,i,final}),H(d^0_{t,j})^2)}{2\alpha_a^2}$$

DB=Bhattacharyya distance. H(.)=RGB color histogram of a text region.

$d^*_{t-1,i,final}$=region of trajectory $T^*_{t-1,i}$ in frame t when last time $T^*_{t-1,i} canbevisible$

Location Similarity is

$$S_l\left(T^*_{t-1,i},d^0_{t,j}\right)=exp\left(\frac{DE\left(l^*_{t-1,i},l^0_{i,j}\right)^2}{2\alpha_i^2}\right)$$

where DE(.)=Euclidean distance l^*_{t-1}=predicted location of a trajectory T^*_{t-1} in frame t. $l^0_{i,j}$=location of detection $d^0_{t,j} inframet$

5.3 Trajectory I and T

if the detections of three consecutive frames are matched,then only new trajectory will be generated. if none or one detection matched then successive frame will be considered as noise.if trajectories are not matched with detections then instead of trajectory is terminated trajectory of another frame is composed by all remaining trajectory.

Chapter 6

Tracking Based Text Detection

In tracking based text detection there are two main steps they are as follows: 1) Detection Revising by Tracking 2) Trajectory Updating

6.1 Detection Revising by Tracking

In tracking based text detection only few detections are considered as noises in trajectory initialization.This step is for constructing detection results.how to recall the missing detections?

- If similarity between the detection and trajectory of last frame is very low then variations of background are complex.
- Appearance similarities of detection result considered as noise then text regions of trajectories in the last 10 frames are calculated.
- The detection is considered as correct only if one of the similarities is larger than threshold which is already set(0.7).
- if detections in frame t matches none of the trajectory from t-1 then color and contour similarity between predicted position and trajectory position in previous 10 frames is calculated.

6.2 Trajectory Updating

Trajectory updating is very simple step.The original detection are not used instead of that revised detections are used in trajectories of frame t.then retrieved detections are updated and added into the corresponding trajectory.

Chapter 7

Text Recognition

Recognition is the process of segmenting.Recognition is constructed using following steps:

- Temporal over-segmentation
- Merging
- Boundary Refinement
- voting

it uses optimal character recognition tool (tesseract-OCR).Extracted text can be given to OCR as input.Then it recognize the text.Trajectory over-segmentation and merging,and multi-frame integration are two major issues of this method.

7.1 Over-Segmentation and Merging

In this step first,each trajectory is temporally and independently over-segmented.The text which has high confidence is considered as reference text.and the frame with reference text is reference frame.Text in other frames is noise text.over-segmentation is better than under-segmentation.because in under-segmentation ID switch is always occurs.

Agglomerative hierarchical clustering algorithm is used for merging process.It selects the two data points and merge them into one single cluster.

7.2 Integration:

In multi-frame integration there are three main steps.

- TBG Identification.
- TBG filteration.
- TBG Integration.

In text block group identification blocks text with same are identified by considering the location.In Text block group filteration blur text is filtered that can bring bad effect to result.In integration to get clean and clear text average and minimum integration an be done.

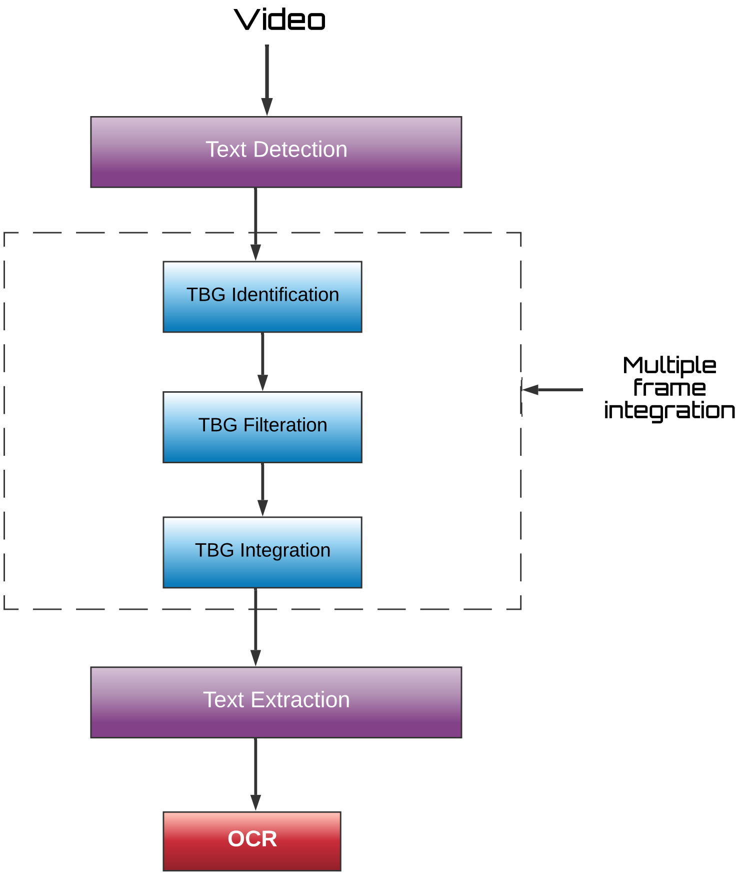


Figure 7.1: Multiframe Integration

Chapter 8

Algorithm

8.1 Agglomerative hierarchical clustering:

Let X = x1, x2, x3, ..., xn be the set of data points.

- Step 1: Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
- Step 2:Find the least distance pair of clusters in the current clustering, said pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.
- Step 3:Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].
- Step 4:Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).
- Step 5:If all the data points are in one cluster then stop, else repeat from step 2).

Chapter 9

CONCLUSIONS

Text Mining is the sub-part of data mining.nowdays increasing growth of online social media and smart phones leads to accumulation of large amount of data.hence,Text in videos is very important for understanding and retrieving multimedia.In the process of text extraction there are three main phases.Tracking,Text Detection,Text Recognition.

In text tracking novel tracking by detection approach is used.it has appearance model for detection and motion model for tracking text.It is totally based on bayes theorem.it finds most probable result.There are various challenges for text detection.these are complex background,similar colors of background and foreground.

Bibliography

[1] Shu Tian,Xu-cheng Yin, *"A Unified framework for tracking based text detection and recognition from web videos,"*, IEEE transactions on pattern analysis and machine intelligence,VOL.,NO.,

[2] x.-C.Yin,Z.-Y.Zuo,S.Tian, and C.-L.Liu, *"Text Detection,tracking and recognition in video:A comprehensive survey,"*, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL.25 , NO.6 2016.

[3] Y.F.Pan,X.Hou,and C.L.Liu, *A hybrid approach to detect and localize texts in natural scene images,*, IEEE Transactions on Image processing,vol.20,no.3 March 2011.

[4] Mr. Mule S. S. Mr. Holambe S. N., *"Detecting Text in Natural Scenes with Connected Component Clustering and Nontext Filtering,"*, International Research Journal of Engineering and Technology (IRJET) Sep 2016.

[5] Pooja Dr. B. R. Ambedker National Institute of Technology, Jalandhar, India

Renu Dhir Dr. B. R. Ambedker National Institute of Technology, Jalandhar, India *"Video Text Extraction and Recognition: A Survey ,"*, IEEE WiSPNET 2016 conference.

[6] Rongrong Wang, Wanjun Jin, Lide Wu Dept. of Computer Science and Engineering, Fudan University, Shanghai, P.R.China (200433) rrwang, jwj, ldwu@fudan *"A Novel Video Caption Detection Approach Using Multi-Frame Integration ,"*,

[7] Boris Epshtein , Eyal Ofek , Yonatan Wexler *"Detecting Text in Natural Scenes with Stroke Width Transform ,"*,

[8] Xiangrong Chen1 Alan L. Yuille1,2 Departments of Statistics1, Psychology2, University of California, Los Angeles, Los Angeles, CA 90095 *"Detecting and Reading Text in Natural Scenes "*,,

[9] Yugandhara Bapurao Dasri1, Bhagyashree Vyankatrao Barde 2, Nalwade Prakash Shivajirao 3, Anant Madhavrao Bainwad *"Text Mining Framework, Methods and Techniques "*,,

[10] Vivek Dhanapal Sapate *"A Survey: Text Extraction from Images and Video"*,, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016

Personal Information

Name: *Harshada Tukaram Deokar*

Date Of Birth: *01 November 1998*

Permanant Address: *At/Post Mohi,Tal-Man,Dist-Satara*

E-Mail: *harshadatd1998@gmail.com*

Mobile: *7038194808*

LinkedIn: *https://www.linkedin.com/in/harshada-deokar-372a02183/*