ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

A Novel Based Text Extraction, Recognition from Digital E-Videos

Ms.Suwarna Baheti[1], Prof.Kalpana S.Thakre[2]

M.E. Student, Department of Information Technology, Sinhgad College of Engineering, Pune, India[1]

Associate Professor, Department of Information Technology, Sinhgad College of Engineering, Pune, India[2]

ABSTRACT: This paper represents text extracting information (represented by data) from a sequence of images (video) is the main objective of Video segmentation. In order to extract and search important information from a huge amount of video data, we are focusing on extraction of texts from video. However variations of the text due to differences in text style, font, size, orientation, alignment as well as low image contrast and complex background make the problem of automatic text extraction extremely difficult and challenging job. A large number of techniques have been proposed to address this problem and the purpose of this paper is to design algorithms for each phase of extracting text from a video using java libraries and classes. Here first we frame the input video into stream of images using the Java Media Framework (JMF) with the input being a real time or a video from the database and considering the connected component analysis form. We apply preprocessing algorithms to convert the shot-frame to gray scale and remove the disturbances like superimposed lines over the text, discontinuity removal, and dot removal. Then we continue with the algorithms for localization, segmentation, tracking and recognition.

Keywords**:** Image Processing, Text Extraction, Text Recognition, Localization, Binarization, Segmentation, Text in video.

I. INTRODUCTION

A Digital video format has a very specific structure. It is a sequence of images with hierarchical structure: from single frames, shots, scenes, and episodes to acts [6]. Most of the image indexing techniques can be adapted for video files. However, because of the huge amount of frames in video files, reducing processing time requires the use of several further techniques like scene detection, or key frame extraction [5], [7], [8]. It is a challenge also because of the low resolution of video files (in comparison with document images), the loss of contrast which is a result of compression (e.g. MPEG) and the complex background [7], [8].

Video text can be classified into two broad categories: Graphic text and Scene text. Graphic text or text overlay is the video text added mechanically by video editors such as caption text or superimposed text. Examples include the news/sports video caption, movie credits etc. Scene texts are the video texts embedded in the real-world objects or scenes. Examples include street name, car license number, and the number/name on the back of a soccer player, E-videos, writing on signs or billboards, text on the sides of trucks or even writing on tee-shirts. These videos contain texts, including scrolling texts or caption text made by artificial overlaying after recording and scene text embedded in backgrounds. Text embedded in images contains large qualities of useful information. Since words have well-defined and unambiguous meanings, text extracted from video clips can provide meaningful keywords which can reflect the rough content of video. These keywords can be used for indexing and summarizing the content of video clip [1].

Text recognition in images and video sequences is a field of research in pattern recognition, artificial intelligence and machine vision, usually called “Video OCR”, which stands for Video Optical Character Recognition. Video OCR attempts to create a computer system which automatically detects, extracts and understands what is meant by the text embedded in the images and video frames.

There are four steps to perform video text recognition

1. Detection: Detect the presence of text, which should answer to “Is there a text string in the current frame?”
2. Localization: Localize the region of text, which should answer to “Where is the text string in the current frame?”
3. Extraction: - Extract the text, generally this step is accompanied with an enhancement processing.
4. Recognition: - Recognize the text, which should answer to “What does this text string say?” generally this step is accompanied with some binarization and/or segmentation pre-processing step.

Difficulties of this can be classified in following main categories:-

i. Background and text may be ambiguous.
ii. Text color may change- text can have arbitrary and non-uniform color.
iii. Background and text are sometimes reversed.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

iv. Text may move.
v. Unknown text size, position, orientation and layout- captions lack the structure usually associated with documents.
vi. Unconstrained background- the background can have colors similar to the text color. The Background may include streaks that appear very similar to character strokes.
vii. Color bleeding-lossy video compression may cause colors to run together.
viii. Low contrast- low bit rate video compression can cause loss of contrast between character strokes and the background.

II. WHY USE VIDEOS IN E-VIDEOS

T People have been using video clips for years; it's just now they are more accessible on computer. Here are some ideas from E-learning Networks Community Forum and E-tools 'n' Tips for Educators community forum on how you can take advantage of videos in e-learning called as E-videos:

i. E-videos are good for illustration purposes. They are supported by text and interactions.
ii. A 20 second E-video can easily replace a full page of text, especially when trying to explain a detailed process or activity (i.e. "How to").
iii. Use of excursions such as e-video the excursion and edit into a movie of less than 5 minutes. Get students to view the e-video and reflect on what happened during the day, what were the highpoints etc.
iv. To show others what the students are doing so that the community can feel part of the learning environment.
v. When you create your our own movies as a group with students as opposed to viewing 'other films', it enhances interactions and personalizes within the groups.
vi. Use to highlight, freeze-frame or slow-motion an event can be of incredible benefit e.g. playing a musical instrument or performing a craft skill or studying a sports sequence.
vii. Pictures in a textbook can be quite boring but through e-video using real scenario can make whole situation so much more engaging.
viii. A short e-video at the start of a lesson helps as an attention grabber to get the students thinking and focused.
ix. To simplify the language whether as subtitles, on-screen labels or voice-overs is of immense benefit to the language derived learner.
x. Learners of varying age groups enjoy and sustain their enthusiasm for learning with computers by first learning about using digital cameras, working as a team and planning a digital story that would have impact in their community.
xi. E-videos can be made by small groups where large classes may not be permitted due to space constraints or Health and Safety issues.
xii. E-videos stored on-line allow students (especially those who are slow learners) to review a topic several times.

III.FRAMEWORK OF TEXT EXTRACTION, RECOGNITION IN E-VIDEOS

In this section, to have a clear overview of these methods, we will only focus on describing the frameworks and the main procedures of the methods.

A] Text detection: Previous text detection methods in complex background can be classified into bottom-up, heuristic top-down methods and machine learning based top-down methods.

A.1] Bottom-up methods-This does not really detect where the text is. The methods directly segment images into regions and then group "character" regions into words.

A.2]Heuristic top-down methods-The first part of the algorithms aims at detecting text regions in images and the second part can be regarded as applying a bottom-up method on a local image.

Machine learning based top-down methods-This system extracted derivative features from fixed-size blocks of pixels and classified the feature vectors into text or non-text.

B] Text recognition: Since commercial OCR engines achieve high recognition performance when processing black and white images at high resolution, almost all the methods in the literature that addressed the issue of text recognition in complex images and videos employed an OCR system to finally recognize characters. To extend the recognition capability of the OCR for image and video text, the main research efforts focus on text segmentation and enhancement.

C] Text segmentation- Methods are performed on the extracted text regions to remove the background surrounding text characters. These methods usually assume that the gray scale distribution is bimodal and that characters a priori correspond to either the white part or the black part. Great efforts are thus devoted to performing better binarization. To

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

eliminate the non-character regions in each binary image, a simple connected component analysis step is employed by setting constraints on size, height and width ratio and so on. However, these methods are unable to filter out background regions with similar gray scale values to the characters.

D] Text enhancement - If the character gray scale value is known, text enhancement methods can help the binarization process. A method for enhancing text in images exploits the characteristic that text characters consist of many stripe structures. The enhancement is performed on text images, which are blocks of image containing the same text string detected and tracked in consecutive video frames.

IV.STEPS TO RECOGNIZE TEXT FROM E-VIDEO

Text information extraction process is usually divided into several steps. The researchers used different names ambiguously and interchangeably. In this, we introduce an unsupervised method to detect and localize text objects in images and video frames. This method is based on a novel pictorial structure based text model and three new character features. In the proposed text model, each character is a part and every two neighboring parts are connected by a link. For every part in the model, we use the presented character features to compute character energy, which can reflect the inherent properties of characters and indicate the probability that a candidate part in the model is a character. For every link in the model, we use the spatial relationship and property similarity between neighboring characters to compute link energy, which indicates the probability that two connected candidate parts are both characters.

The advantages of the proposed method are: (i) the characteristics of character and the structure of text object are described by the parts. Therefore, the proposed method can capture the properties of characters and text objects simultaneously and combine them efficiently; (ii) the presented three new character features are computed based on the inherent properties of character. Therefore, the proposed method is robust to the size, font, color, and orientation of text and can discriminate text objects from other objects efficiently. The steps of the proposed method as follows: (i) Initialize candidate text models by localizing the candidate parts and connections in a given frame image. (ii) Compute character energy for each part based on the character properties. (iii) Compute link energy for each connection based on the text properties. (iv) Compute text unit energy and use minimum spanning tree to generate final text models as shown in figure 1.

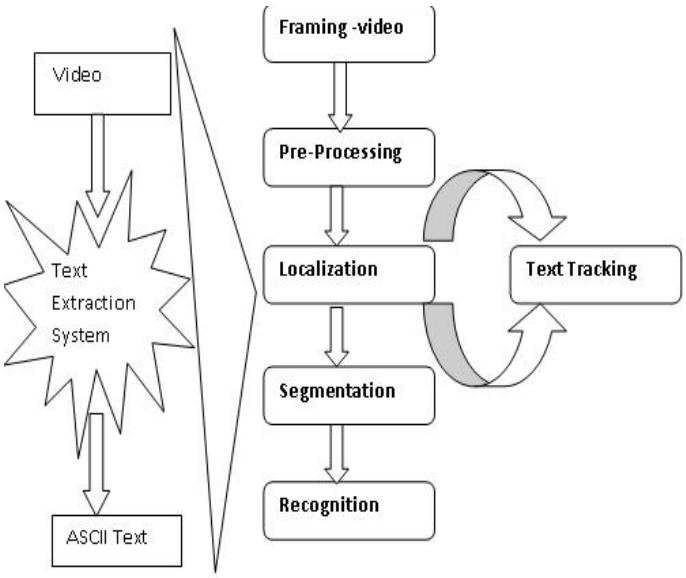


Figure 1. Text Recognition Steps

A] Framing of Video

We have used the Java Media Framework (JMF) to capture the media content and frame the video. JMF is a framework for handling streaming media in Java programs. JMF is an optional package of Java 2 standard platform. JMF provides a unified architecture and messaging protocol for managing the acquisition, processing and delivery of time-based media.

JMF enables Java programs to:

i. Present (playback) multimedia contents.
ii. Do real-time streaming of media over the internet.
iii. Process media such as changing media format, adding special effects.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

iv. Store media into a file.
v. JMF provides a platform-neutral framework for handling multimedia

The input to this stage was a video containing text. The video was then framed into images using JMF at the rate of 1 frame per second. This rate could be increased or decreased depending on the speed of the video i.e. on the basis of fps (frames per second). The images were scaled to a resolution of 280x90 and were saved at the specified location on the hard disk drive.

For example: If the video was of 30 seconds, then 30 frames (images) of the video would be scaled to a size of 28x90 and saved which were then given as an input to the next stage [7].

B] Pre Processing

A scaled image was the input which was then converted into a gray scaled image. This image formed the first stage of the pre-processing part. This was carried out by considering the RGB color contents(R: 11%, G: 56%, B: 33%) of each pixel of the image and converting them to gray scale. The conversion of a colored image to a gray scaled image was done for easier recognition of the text appearing in the images as after gray scaling, the image was converted to a black and white image containing black text with a higher contrast on white background.

The second stage of pre-processing is lines removal. A video can contain noise which is either a horizontal fluctuation (a horizontal line throughout the screen) or a vertical fluctuation (a vertical line throughout the screen). Thus, for successful recognition of the text appearing in the frames, there is a necessity that these horizontal and vertical fluctuations should be removed. This was carried out by clearing all pixels (changing pixel color from black to white) located on all lines which appeared horizontally and vertically across the screen because of the fluctuations which may have occurred in the video. This stage did not make any changes to the image if the video frame did not contain any horizontal and vertical fluctuations [6].

The third stage of pre-processing is discontinuities removals that were created in the second stage of pre-processing. As explained above, if the video contained any fluctuations, then those fluctuations were removed in the lines removal stage. If the horizontal and vertical fluctuations occurred exactly where the text was present, then it created discontinuities between the texts appearing in the video frame which made the recognition of the text very difficult. This was carried out by scanning each pixel from top left to bottom right and taking into consideration each pixel and all its neighboring pixels. If a pixel under consideration was white, and all the neighboring pixels were black, then that corresponding pixel was set as black because all the black neighboring pixels indicated that the pixel under consideration was cleared at the lines removal stage because of the fluctuations [3].

The final output of pre-processing stage is wherein the remaining disturbances like noise are eliminated. This was carried out again by scanning each pixel from top left to bottom right and taking into consideration each pixel and all its neighboring pixels. If a pixel under consideration was black and all the neighboring pixels were white, then that corresponding pixel was set as black because all the black neighboring pixels indicated that the pixel under consideration was some unwanted dot (noise) [2].

C] Detection and Localization

In the text detection stage, since there was no prior information on whether or not the input image contains any text, the existence or non-existence of text in the image must be determined. However, in the case of E-video, the number of frames containing text is much smaller than the number of frames without text. The text detection stage seeks to detect the presence of text in a given image. Selected a frame containing text from shots elected by video framing, very low threshold values were needed for scene change detection because the portion occupied by a text region relative to the whole image was usually small. This approach is very sensitive to scene change detection. This can be a simple and efficient solution for video indexing applications that only need key words from video clips, rather than the entire text. The localization stage included localizing the text in the image after detection. In other words, the text present in the frame was tracked by identifying boxes or regions of similar pixel intensity values and returning them to the next stage for further processing. This stage used Region Based Methods for text localization. Region based methods use the properties of the color or gray scale in a text region or their differences with the corresponding properties of the background. [2], [5].

D] Segmentation

After the text was localized, the text segmentation step deals with the separation of the text pixels from the background pixels. The output of this step is a binary image where black text characters appear on a white background. This stage included extraction of actual text regions by dividing pixels with similar properties into contours or segments and discarding the redundant portions of frame [2].

E] Recognition

This stage included actual recognition of extracted characters by combining various features extracted in previous stages to give actual text with the help of a supervised neural network. In this stage, the output of the segmentation stage is considered and the characters contained in the image were compared with the pre-defined neural network training set and depending on the value of the character appearing in the image, the character representing the closest training set value was displayed as recognized character [2], [4].

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

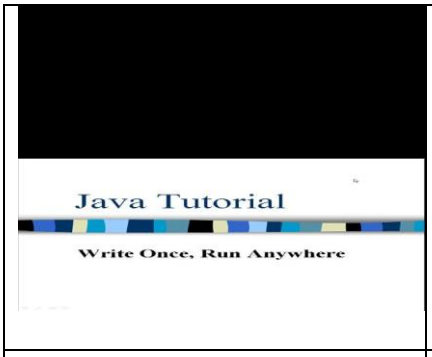
International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

V. EXPERIMENTAL RESULTS

The goal of the experiment is to analyses the influence that the quality of text included in the video has on the success rate of text information extraction (recall and precision in groups of Extraction probability); namely usefulness of text information extraction in video indexing and information retrieval.

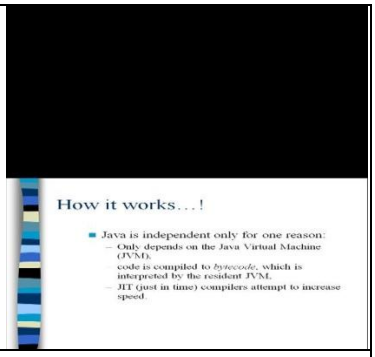
The recall and precision have been measured for each mentioned above group of text quality separately in order to capture the influence of text quality on the success of text information extraction. The processing time has been measured together to obtain the result for an average frame. The errors of run have been reported. Finally, the results have been analyzed and evaluated in the context of usefulness in digital E-video information retrieval. Different word recognition rate has been measured for each specified category separately in order to verify the usefulness of text information extraction methods for providing video indexing. The usefulness has been specified by the obtained results: amount of different words correctly recognized per category in video.

Example. Input is Digital E-video of java which has explained by Instructors'

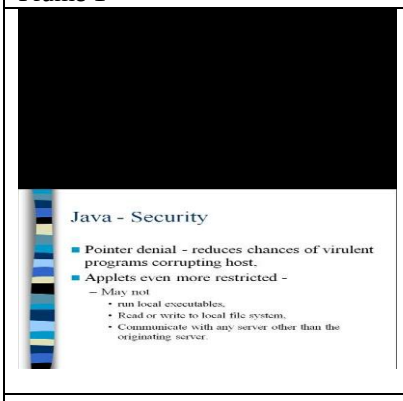


Frame 1

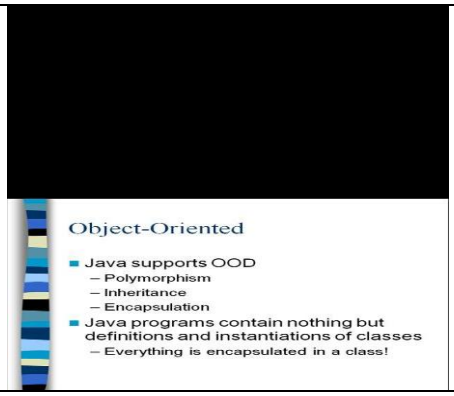
Frame 2



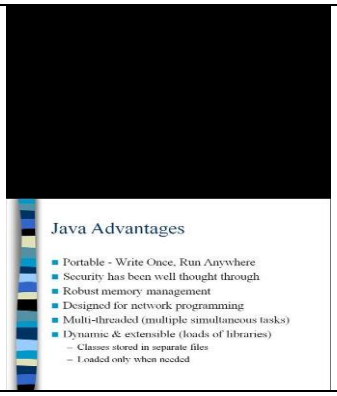
Frame 3



Frame 4



Frame 5



Frame 6

	TC	RL-RT	RT-NRL	NRT-RL	Precision	Recall
Frame1	33	33	00	00	01	01
Frame2	186	179	06	02	0.96	0.98
Frame3	180	170	10	05	0.94	0.97
Frame4	197	173	24	08	0.88	0.95
Frame5	166	164	01	02	0.99	0.98
Frame6	250	218	30	05	0.87	0.97

Table 1. Frame Database Results

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 5, July 2013

TC: Total no. of Characters **RL:** Relevant**; RT:** Retrieved **NRL:** Non Relevant; **NRT:** Not Retrieved

Therefore, Proposed Approach is
Precision = 94% Recall = 97.5%
Different words Recognition rate = 95%
Computational Time = approximately 2-3 seconds

VII. CONCLUSION

In this paper, we propose a novel framework simultaneously considering the instructors' lecture videos and proposes novel techniques for video segmenting, multimedia knowledge including techniques for discovering perceptual and semantic knowledge for e-videos. The interaction can be discovered when video be recorded, edited, and playback. For material creation, we reveal a common problem and propose a solution. A recording model is also given to enhance instructors' interactions into the material.

Recognition of keyword from a video sequence is still one of the most challenging problems in educational material because video is of low quality and the frame images are small. We have proposed a simple and efficient technique to detect and recognize keyword from a video sequence and these are two major challenges: (detection and segmentation). Detection is not difficult but accurate and correct material from video is hard to acquire.We study the performance of algorithm that the algorithm is sensitive to the number of sequences and gives good result which understands to the new user.

REFERENCES

[1] Julinda Gllavata,"Extracting Textual Information from Images and Videos for Automatic Content Based Annotation and Retrieval."

[2] Oxford Clarendom, "An Edge-based Approach on Electricity and Magnetism," IEEE,3rd ed., vol 2,1892,pp.68-73.

[3] "A Robust Algorithm for Text Extraction in Color Video" (This Paper appears in Multimedia and Expo. 2000, ICME 2000, 2000 IEEE International Conference on Issue Date:2000)

[4] K.Jung, K.I.Kim and A.K.Jain, "Text Information Extraction in Images and Videos: A Survey Pattern Recognition Letters," 27: 977-997,2004.

[5] R.Lienhart and A.Wernicke, "Localizing and Segmeniting Text in Images and Videos," Transactions on Circuits and Systems for Video Technology, 12(4):256-268,2002.

[6] N.Efford , "Digital Image Processing: a Practical Introduction using Java," Addison Wesley,2000.

[7] Yeo, B., Liu, B. (1995), "Rapid scene analysis on compressed video," IEEE Transactions on Circuits & Systems for Video Technology, 533-44.

[8] Zhang, H., Kankanhalli A., and Smoliar, W. (1993), "Automatic partitioning of full-motion video," Multimedia Systems, 10-28.

[9] Chitra Dorai, Oria, V., Neelavalli, V., "Structuralizing educational videos based on presentation content," Image Processing, 2003 International Conference on, Vol. 2, pp-1029-32, 14-17 Sept. 2003.

[10] A. F. Smeaton, "Indexing , Browsing , and Searching of Digital Video and Digital Audio Information," *Audio*, pp. 93-110, 2000.

[11] J. Zhang and R. Kasturi, "Extraction of Text Objects in Video Documents: Recent Progress," *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, pp. 5-17, Sep. 2008.

[12] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption," *Multimedia Systems*, vol. 7, no. 5, pp. 385-395, 1999.

[13] H. Li and D. Doermann, "Text Enhancement in Digital Video Using Multiple Frame Integration," *Methodology*, pp. 1-12, 1999.

[14] D. Chen and J. Odobez, "Video text recognition using sequential Monte Carlo and error voting methods," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1386-1403, Jul. 2005.

[15] H. Z. Xiaodong Huang and M. Huadong "A New Video Text Extraction Approach," *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 650-653, 2009.

[16] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo, "Extracting Semantic Information from News and Sport Video," ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat. No.01EX480), pp. 4-11, 2001.

[17] K. Jung, "Text Information Extraction in Images and Video: a Survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977-997, May 2004.

[18] Chen D., J. Luettin, K. Shearer, "A Survey of Text Detection and Recognition in Images and Videos", Institute Dalle Molled' Intelligence Perceptive (IDIAP) Research Report, IDIAP-RR 00-38, 2000

[19] Jung K., K.I. Kim, and A.K. Jain, "Text Information Extraction in Images and Video: A Survey", Pattern Recognition, pp. 977-997, 2004