Video Retrieval Based on Textual Queries

C. V. Jawahar, Balakrishna Chennupati, Balamanohar Paluri, Nataraj Jammalamadaka
Center for Visual Information Technology,
International Institute of Information Technology,
Gachibowli, Hyderabad - 500 019
jawahar@iiit.net

Abstract

With the advancement of multimedia, digital video creation has become very common. There is enormous information present in the video, necessitating search techniques for specific content. In this paper, we present 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 the text within the video. The choice of OCRs brings in many constraints on the language and the font that they can take care of. We hence propose an approach that enables matching at the image-level and therby avoiding an OCR. Videos containing the query string are retrieved from a video database and sorted based on the relevance. Results are shown from video collections in English,Hindi and Telugu.

1 Introduction

Large amount of multimedia information is getting generated in various domains. The need for efficient techniques for accessing the relevant information from these multimedia databases is on the rise. Content based information retrieval techniques are the bare minimum to access the relevant information in these collections. Search and retrieval techniques have reached a maturity for textual data resulting in powerful search engines. Active research has been seen in the area of image retrieval and video retrieval during the last decade [6, 7]. Large video collections, which were thought of impossible at one stage, are becoming increasingly common due to the advancement of storage technologies.

Conventional approaches for video indexing are based on characterizing the video content using a set of computational features [2]. These techniques often do not exploit the commonly available high-level information conveyed by textual captions. Retrieval of relevant videos from large video databases has immense applications in various domains. Retrieval from broadcast news video database is vital for effective information access. The presence of textual captions and audio, complementing the appearance information of video frames, enables building automated retrieval systems. The information about the song and performer, present in the form of text captions enables searching in albums of an artist,a genre of songs. These techniques can also find uses in digital libraries of video clips of lectures for distance learning applications.

There are two important issues in Content-Based Video Access: (a) A representational scheme for the content (b) A Human friendly query/interface. In this paper, we propose a novel approach to video retrieval which is significantly better than most other reported techniques on both these fronts. Our method is highly suited for videos containing textual information, like video clips from news broadcasts.

Manual annotation of frames in videos has been useful for indexing for quite a long time. Though this may be very precise, it is highly cumbersome to generate. Methods were also proposed for analyzing the text regions as blocks, lines, words and characters, and finally recognizing the characters using OCR systems to output the text strings contained in the image. These text strings are then used as keywords [4] for indexing. Other techniques for content based video retrieval using visual features, have also been partially successful in retrieval based on semantic content [10]. Our efforts are focused on using the textual information for indexing. We describe a system that extracts the textual information from the frames of the video. Without deciphering the text using an OCR, we index the video using the images of the text. To be more precise we index the video using a set of features instead of a well defined alphabet. Given a textual query, therefore we can search for the related videos (where the word is present) and rank them based on attributes like size of the text and duration of its appearance.

This paper is organized as follows: Section 2 explains the problem, previous work in the area and introduces our approach. Section 3 explains our algorithm in detail and presents the block diagram of the system. Section 4 deals with the feature extraction and matching modules and explains the process in a greater detail. The results of the scheme are reported in the Section 5. Finally conclusions and future work are presented in Section 6.

2 Problem Setting

2.1 Video Retrieval Schemes

Several approaches have been reported for indexing and retrieval of video collections. They model spatial and temporal characteristics of the video for representation of the video-content. In spatial domain, feature vectors are computed from different parts of the frames and their relationship is encoded as a descriptor. The temporal analysis partitions the video into basic elements like frames, shots, scenes or video-segments. Each of the video segments are then characterized by the appearance and dynamism of the video content. It is often assumed that features like, histograms, moments, texture and motion vectors, can describe the information content of the video clip.

In a database of videos, one can query for relevant videos with example images, as is popular for content based image retrieval. Extending this approach has the limitation that it does not utilize the motion information of the video and employs only the appearance information. Moreover, finding example video clips for the concept of interest can be quite complex for many applications. Textual query is a promising approach for querying in video databases, since it offers a more natural interface.

There are many powerful clues hidden in the video clips. Audio and the textual content in videos can be of immense use in indexing. Textual information is present as captions appearing on the video or printed/handwritten text in the scene. If we can detect and recognize the textual content, it can be effectively used for characterizing the video clips, as a complementary measure to the visual appearance-based features. The audio content can fill up the missing information that has not been displayed in the caption. Speech recognition applications facilitate the ability to use the audio content for indexing. There are attempts in extracting the textual information present in the video using OCRs for video indexing [4]. Textual information, especially in news videos has the essential information needed for retrieval. However, these attempts are not good enough for many of the video broadcasts. For broadcasts in Indian languages, we do not have OCRs which can convert the text image into machine processible text.

2.2 Text Detection in Video Frames

An important step in characterizing video clips based on the textual content is the robust detection of the textual blocks in images/frames. Binarization techniques, which use global, local, or adaptive thresholding, are the simplest methods for text localization in images. These methods are widely used for document image segmentation. Text detection in videos and natural images needs more advanced techniques. Methods which detect text in video use either region property in the spatial domain or the textural characteristics [5].

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. These approaches work in a bottom-up fashion by identifying sub-structures, such as connected components or edges, and then merging these substructures to mark bounding boxes of the text. A geometrical analysis is needed to merge the text components using the spatial arrangement of the components. Edge-based methods focus on the high contrast between the text and the background. The edges of the text boundary are identified and merged. Then several heuristics are used to filter out the non-text regions. Texture-based methods use the observation that text in images have distinct textural properties that distinguish them from the background. Techniques based on Gabor filters, Wavelet, FFT, spatial variance, etc. have been used to detect the textural properties of a text region in an image [1]. Many of these methods are also suitable for processing in the compressed domain [11].

Due to the large number of possible variations in text in different types of applications and the limitations of any single approach to deal with such variations, many researchers have developed hybrid approaches. Some fused the connected component-based approach with the texture-based approach to overcome the drawbacks of the individual approaches.

2.3 Proposed Method

Text within an image is of particular interest for indexing and retrieval of video because of its capability to describe the contents of an image, and its relationship to the semantic information. It also enables applications such as keyword based search in multimedia databases. In this work, we propose to use the *text-images* for video indexing and retrieval. This method works without the dependency on the availability of an OCR.

As a first step, we extract text blocks in video frames. Extraction of text information involves detection, localization, enhancement and recognition of the textual content in the video frames [5]. Our method involves a frame by frame processing on the entire video for locating textual blocks. Each frame is divided into regions of size $N \times N$, where N depends on the size of the frame. For our experiments, we divided a 320 $\times$ 240 frame into 25 parts. These regions are separated into different classes using Multi-level Thresholding. These classes of pixels are then arranged into planes based on the pixel and location similarities. On each of these planes, we perform connected component analysis followed by XY-cut [3] to detect and locate the text. We use these textual regions for indexing and retrieval of the video. We match words at the image-level, without any explicit textual representation for providing the access to the video database. Matching and retrieval are described in a greater detail in the next sections.

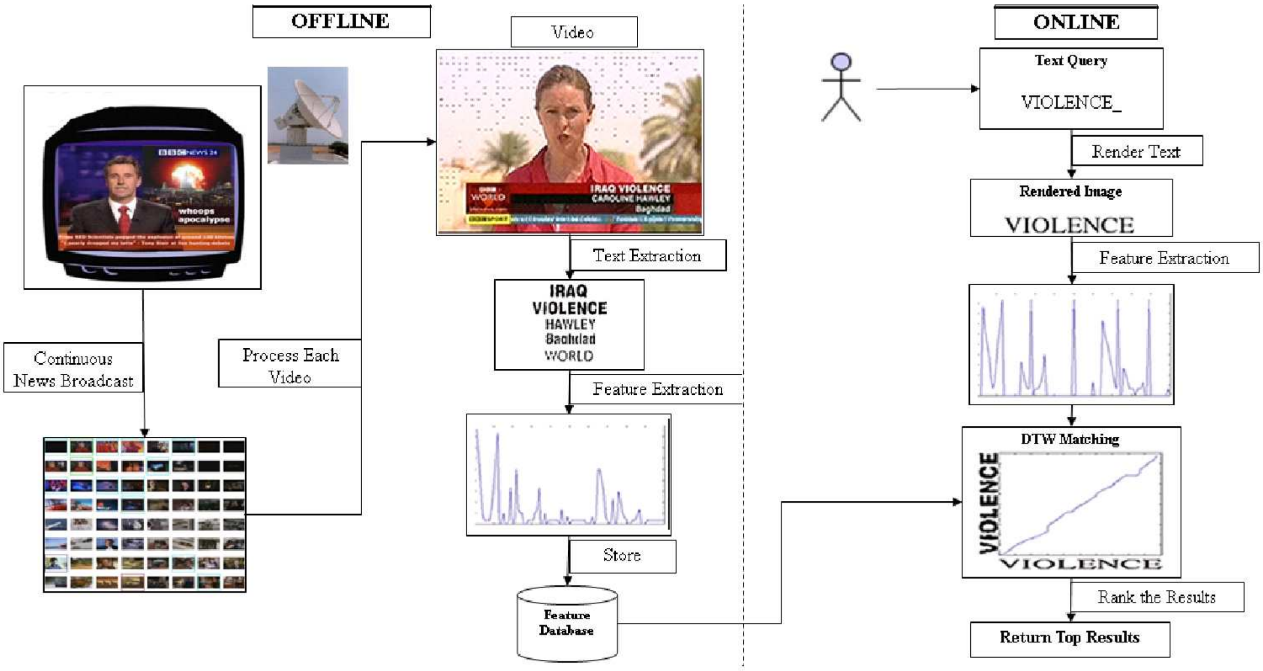

Figure 1: A conceptual diagram of the text-image-based video retrieval

3 Textual Query for Video Retrieval

An advanced video retrieval solution could identify the text present in the video, recognize the text, compute the similarity between the query string and pre-indexed textual information present in the video. However, success of this technique depends on two important aspects: (a) Quality of the input video (b) Availability of an OCR for robustly recognizing the text images. In many practical situations, we find video clips where the resolution of the broadcast is not enough even for a reasonable robust OCR to work on. Moreover for many of the languages, we do not have OCRs available for decoding the textual content in the frames. Since we do not have OCRs available to work effectively on the text in the video data, we use text images to index the videos.

A conceptual diagram of the video retrieval process we follow is given in Figure 1. This has two phases – online and the offline phase. During the offline phase, broadcast videos in multiple languages are stored in a video database. Cuts and theme changes are identified in each video for the segmentation. These videos are further processed frame by frame for identifying possible text regions as described in the previous section. A selected set of features is extracted from these text regions and stored in a feature database.

During the online phase, a search query is entered through a graphical user interface. This query string is rendered as an image and the corresponding set of features is extracted. These features are same as those employed in the offline process. A matching algorithm then computes the degree of similarity between the features of the search query and those present in the feature database. The results are then ranked based on the similarity measure. Word form variations are handled using a partial matching method, based on Dynamic Time Warping (DTW). In the next section, we describe the feature extraction and word matching in detail.

We have also used scrolling text as a source of information. Scrolling text, unlike the overlaid text captions, is in motion. Initially the scrolling speed of text is computed. Then a long text image of the scrolling content is generated and connected component analysis is performed to find the words. This text image is used for indexing the corresponding Video clip.

In our experimental system, we focus on broadcast news videos. We experimented with recorded videos from channels of three languages: Hindi, Telugu and English. We used a TV tuner card attached to a personal computer and programmed for regular recordings. We recorded the videos at a resolution of 320 × 240. The recording quality of the videos have an effect on the search as poor quality videos affect the text detection module.

Issues related to the efficient indexing in feature databases are not discussed in this paper. With an incremental clustering scheme and an appropriate hash function, this issue can be taken care.

4 Matching at the Image level

This section details the procedure used to retrieve the videos containing the search query. The video clips are indexed according to the textual regions extracted from the videos. Text-images are used instead of text strings for indexing and retrieval except that the notion of the alphabet is replaced by an abstract feature-representation. For each word, we derive a sequence of feature vectors by splitting the word images vertically. For each of the strips, features based on their profiles, structural property etc. are extracted. These features are computed very similar to the method described in [8], where a search engine for content-based access to document image collections in a digital library is presented.

Profile-based features are useful as shape signatures. Upper profile, Lower profile and Transition profiles are employed here as features for representing the words. Background-to-Ink Transition records the number of transitions from an ink pixel to the background pixel for every image column. Upper and Lower Profiles are computed by recording the distance from the top/bottom boundary of the word image to the closest ink pixel for each image column i.e. the distance between the upper/lower boundary of the word image and the first pixel in each column of the image. These features are normalized before using them for characterizing the text-block. Structural features like mean and standard deviation are also used for characterizing images. Mean is used to capture the average number of pixels per column in a word image. Standard deviation is the sum of squared deviation of pixels per column from the mean. It measures the spread of ink pixels in the word image.

The combination of Lower profile, background-foreground transition and standard deviation performed well for most of the images. For English and Telugu videos, we found that transition, Upper, Lower Profiles along with the standard deviation perform satisfactorily. For Hindi, however, upper profiles were not informative because of the *Shirorekha*. We, hence used the mean as an additional feature for Hindi.

4.1 Word matching and retrieval

Given a textual query, we render an image of the text query, extract features and match with the feature sequences stored in the database. We employ a matching algorithm which compares the features of the rendered image and features present in the database. In [9], it is shown that using Dynamic Time Warping (DTW), can be used for finding similarity of images. The DTW is a popular method for matching sequences like strings and speech samples. Hence we use DTW to compare the feature sequences corresponding to images in video frames. In order to handle the size variations, we normalize all the features during computation. Our approach enables searching in videos with text in any language. The feature extraction scheme is also practically invariant to variations like font, size and style.

4.2 Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is a dynamic programming based procedure used to align sets of sequences of feature vectors and compute a similarity measure. DTW computes the optimal alignment between sequences of feature vectors, so that the cumulative distance measure consisting of local distances between aligned sequences of signals is minimized. The mapping is mostly non-linear where the input signals are of different lengths.

Let the features extracted from the word images be $A_1, A_2, A_3, \ldots A_M$ and $B_1, B_2, B_3, \ldots B_N$. Then DTW cost between the two sequences is calculated using the equation:

$$D(i,j) = min \begin{cases} D(i-1,j-1) \\ D(i,j-1) \\ D(i-1,j) \end{cases} + d(i,j) \qquad (1)$$

Where $d(i,j)$ is the cost in aligning the i^{th} element of A with j^{th} element of B and is computed using a simple squared Euclidean distance.

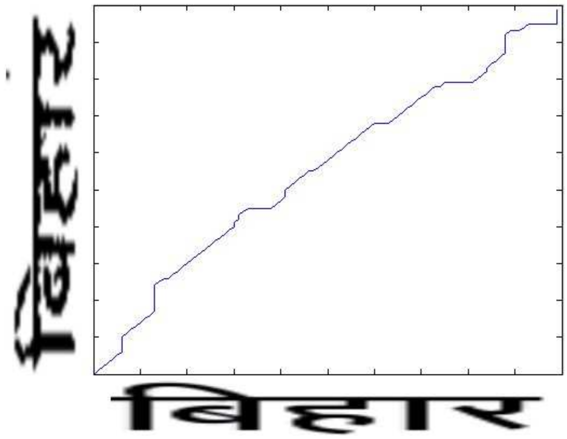

Figure 2: Dynamic Programming based alignment of two sequences of features computed from Hindi words.

Score for matching the two sequences A and B is considered as $D(M,N)$ where M and N are the lengths of the two sequences. The score is then normalized over the width of the image so that it is not biased towards the wider word. The optimal warping path is generated by backtracking the DTW minimal score in the matching space. The length of the optimal warping path is the number of transitions from $D(0,0)$ to $D(M,N)$.

5 Results and Discussions

We have experimented the proposed approach on a collection of broadcast videos. Video clips were collected and processed for detection of Cuts/Segments. Segmentation is done by analyzing the changes in the color/motion statistics corresponding to the text regions over time. Textual regions are detected in these clips. For each of the text regions, the features are extracted and used for matching and indexing. For each of the segments, segment-level information (start and end frames) is stored with the possible characteristic features. Given a textual query, the search module returns the name of the video, location of the text and the duration for which the text is present along with many other attributes of the text block.

For a given textual query, there could be multiple video clips, with the possible occurrence of the word. We need a mechanism for computing the relevance of the video clips to the query, as is done in the conventional information retrieval literature. We rank the videos in such a way that the first result contains more information about the text than the following ones. To quantify the importance of the particular video we have observed that the following factors could be used for computing the relevance of the video clip to the query: (a) Duration for which the text is seen (measured in terms of frames where the text is present) (b) Size of the text (c) Location of the text (d) Characteristics like Scrolling/Static. We have used these measures presently in an adhoc manner for ranking the videos. We are working on defining a comprehensive importance measure for the video clips. For building a well tested ranking scheme, one may need a large database of video clips and human feedbacks on the measures.

Video clips retrieved for the textual queries in English, Hindi and Telugu are shown in Figures 3, 5 and 7. The Figures 4, 6, 8 show the respective instances of the search query in the retrieved frames. For the query "Australia" the top six results given by the system are shown. There are results from different channels. Observe that the system returned results which contained "Australian". The invariance to font sizes can also be observed in the returned results. Also, observe the order of the results, some results are from a news story while some others are from headlines. The ranking methodology used includes the duration of appearance of text and the number of occurrences of the text, the headlines appear for a shorter duration and the news stories for a longer duration. Thus, the news stories are ranked higher.

For the Telugu and Hindi results, we show here the results for the queries "Adhyakshudu" and "Sania" respectively. Here also, one can observe independence of font styles and sizes. For the query "Sania", a result containing "Sonia" has been returned. Since the database has only five instances of "Sania" the nearest matching "Sonia" is returned. This may be attributed to the fact that "Sania" and "Sonia" differ only in their upper profile, which we are not considering as a potential feature in case of Hindi. Similarly in telugu for the query "Adhyakshudu" a result containing "Adhyakshulu" was retrieved as it is the next best match.

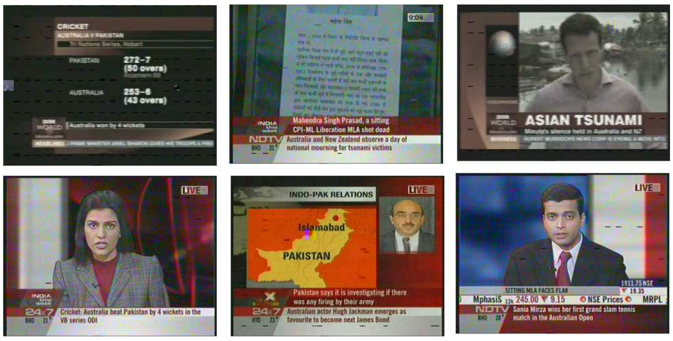

Figure 3: Top 6 Results on English Videos for the query "Australia"

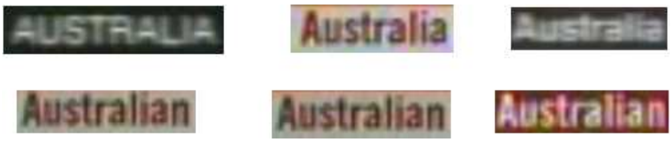

Figure 4: Queried word "Australia" in the respective frames

The system was able to handle variations in font styles and sizes. We have demonstrated the system for English, Hindi and Telugu. However, the system can handle other languages as well.

Our present approach uses textual content for indexing the videos. Here, we are limited by the amount of textual information present in the captions. This approach can be used as a complementary one to the feature based video retrieval schemes. Often, the problem with regional Indian news channels is the absence of textual captions or the presence of a minimal number of captions. We have seen that off-late, many of these channels are giving more importance to the textual content. For some of the Telugu channels, captions were present only for the headlines.

The text detection algorithm is often sensitive to video quality. Also, there is no detailed study reported on the challenges in detecting Indian Language text in videos and natural images. Languages like Hindi, need different approach for text detection since the average length of connected component can be considerably different from that of English. In videos which are noisy, the words are detected along with large amount of false alarms. In videos which have poor resolution, the words are often poorly detected. Another important issue is selection of the best word image from a sequence of its appearances in multiple frames. This becomes more critical in noisy video-broadcasts. We hope, many of these problems will not be there in digital broadcasts and reception.

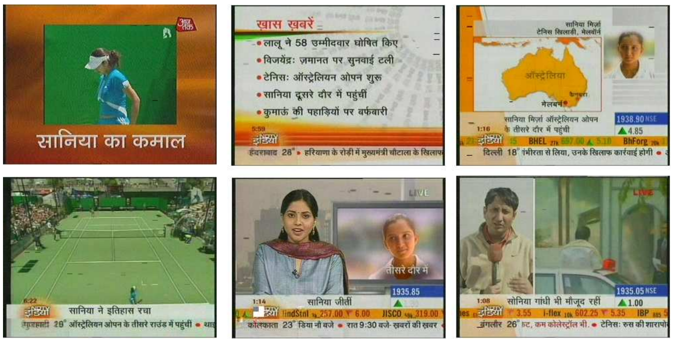

Figure 5: Top 6 Results on Hindi Videos for the query "Sania"

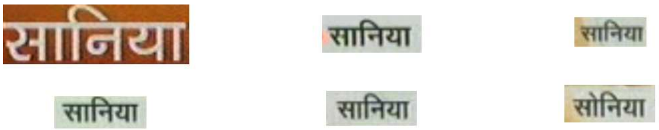

Figure 6: Queried word "Sania" in the respective frames

6 Conclusions and future work

We described an approach for video search based on the text present in the videos using dynamic time warping. We also argued that this approach has wider applicability compared to other techniques including the one based on OCRs. Our future work will focus on improving the accuracy as well as the speed of techniques used here. Accuracy can be improved by using better techniques as well as using a large feature set which discriminates words better from each other. Speed can be improved by optimizing our implementation of the Dynamic Time Warping algorithm as well as looking at related computational techniques to minimize the number of possible matches. We are also exploring an efficient indexing scheme.

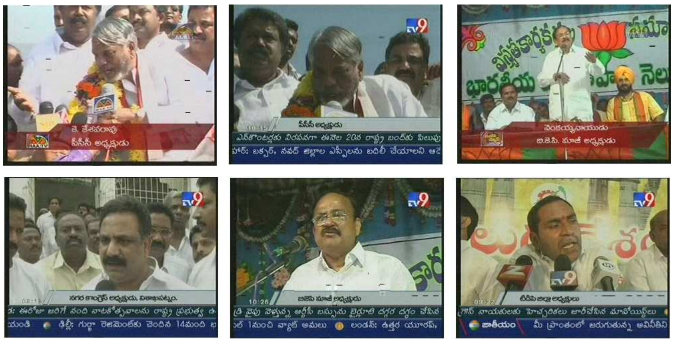

Figure 7: Top 6 Results on Telugu Videos for the query "Adhyakshudu"

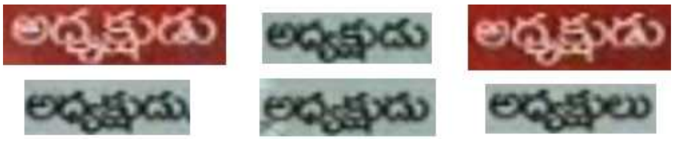

Figure 8: Queried word "Adhyakshudu" in the respective frames

References

[1] A.K.Jain and Bin Yu. Automatic text location in images and video frames. *Pattern Recognition.Vol.3*, 3:2055–2076, Dec 1998.

[2] Andreas Girgensohn John Adcock Matthew Cooper and Lynn Wilcox. Interactive search in large video collections. *Conference on Human Factors in Computing Systems*, April 2005.

[3] Jaekyu Ha, Robert M. Haralick, and Ihsin T. Phillips. Recursive xy-cut using bounding boxes of connected components. *3rd international conference on Document Analysis and Recognition*, pages 952–955, 1995.

[4] T. Sato T. Kanda E. K. Hughes and M. A. Smith. Video ocr for digital news archive. In *Proc. of IEEE Workshop on Content based Access of Image and Video Databases*, pages 52–60, 1998.

[5] Keechul Jung Kwang in Kim and Aanil K.Jain. Text information extraction in images and video: a survey. *Pattern Recognition 37*, pages 977–997, 2004.

[6] Marchand. Content based video retrieval: An overview. http://viper.unige.ch/ marchand/CBVR/.

[7] Wtephane Marchand-Maillet. Content-based video retrieval: An overview. Technical report, University of Geneva, 2001.

[8] C.V.Jawahar Million Meshesha and A. Balasubramanian. Searching in document images. *ICVGIP*, pages 622–628, Dec 2004.

[9] Toni M. Rath and R. Manmatha. Word image matching using dynamic time warping. *Conference on Computer Vision and Pattern Recognition*, pages 521–527, 2003.

[10] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th International Conference on Computer Vision(ICCV), Nice, France*, Oct 2003.

[11] Yu Zhong Hongjiang Zhang and Anil K. Jain. Automatic caption localization in compressed video. *IEEE*, pages 385–392, 2000. IEEE Transactions on Pattern Analysis and Machine Intelligence.