Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2010, Article ID 486487, 18 pages
doi:10.1155/2010/486487

Research Article

Multimodal Indexing of Multilingual News Video

Hiranmay Ghosh,[1] Sunil Kumar Kopparapu,[2] Tanushyam Chattopadhyay,[3] Ashish Khare,[1] Sujal Subhash Wattamwar,[1] Amarendra Gorai,[1] and Meghna Pandharipande[2]

[1] *TCS Innovation Labs Delhi, TCS Towers, 249 D&E Udyog Vihar Phase IV, Gurgaon 122015, India*
[2] *TCS Innovation Labs Mumbai, Yantra Park, Pokhran Road no. 2, Thane West 400601, India*
[3] *TCS Innovation Labs Kolkata, Plot A2, M2-N2 Sector 5, Block GP, Salt Lake Electronics Complex, Kolkata 700091, India*

Correspondence should be addressed to Hiranmay Ghosh, hiranmay.ghosh@tcs.com

Received 16 September 2009; Revised 27 December 2009; Accepted 2 March 2010

Academic Editor: Ling Shao

The problems associated with automatic analysis of news telecasts are more severe in a country like India, where there are many national and regional language channels, besides English. In this paper, we present a framework for multimodal analysis of multilingual news telecasts, which can be augmented with tools and techniques for specific news analytics tasks. Further, we focus on a set of techniques for automatic indexing of the news stories based on keywords spotted in speech as well as on the visuals of contemporary and domain interest. English keywords are derived from RSS feed and converted to Indian language equivalents for detection in speech and on ticker texts. Restricting the keyword list to a manageable number results in drastic improvement in indexing performance. We present illustrative examples and detailed experimental results to substantiate our claim.

1. Introduction

Analysis of public newscast by domestic as well as foreign TV channels for tracking news, national and international views and public opinion is of paramount importance for media analysts in several domains, such as journalism, brand monitoring, law enforcement and internal security. The channels representing different countries, political groups, religious conglomerations, and business interests present different perspectives and viewpoints of the same event. Round the clock monitoring of hundreds of news channels requires unaffordable manpower. Moreover, the news stories of interest may be confined to a narrow slice of the total telecast time and they are often repeated several times on the news channels. Thus, round-the-clock monitoring of the channels is not only a wasteful exercise but is also prone to error because of distractions caused while viewing extraneous telecast and consequent loss of attention. This motivates a system that can automatically analyze, classify, cluster and index the news-stories of interest. In this paper we present a set of visual and audio processing techniques that helps us in achieving this goal.

While there has been significant research in multimodal analysis of news-video for their automated indexing and classification, the commercial applications are yet to mature. Commercial products like BBN Broadcast monitoring system (http://www.bbn.com/products_and_services/bbn_broadcast_monitoring_system/) and Nexidia rich media solution (http://www.nexidia.com/solutions/rich_media) offer speech analytics-based solution for news video indexing and retrieval. None of these solutions can differentiate between news programs from other TV programs and additionally cannot filter out commercials. They index the complete audio-stream and cannot define the story boundaries. Our work is motivated towards creation of a usable solution that uses multimodal cues to achieve a more effective news video analytics service. We put special emphasis on Indian broadcasts, which are primarily in English, Hindi (Indian national language), and several other regional languages.

We present a framework for multimodal analysis of multilingual news telecasts, which can be augmented with tools and techniques for specific news analytics tasks, namely delimiting programs, commercial removal, story boundary

detection and indexing of news stories. While there has been significant research in tools for each of the tasks, an overall framework for news telecast analysis has not yet been proposed in literature. Moreover, automated analysis of Indian language telecasts raises some unique challenges. Unlike most of the channels in the western world, Indian channels do not broadcast "closed captioned text", which could be gainfully employed to index the broadcast stream. Thus, we need to rely completely on audio-visual processing of the broadcast channels. Our basic approach is to index the news stories with relevant keywords discovered in speech and in form of "ticker text" on the visuals. While there are several speech processing and OCR techniques, we face significant challenges in using them for processing Indian telecasts. The major impediments are (a) low resolution (768×576) of the visual frames and (b) significant noise introduced in the analog cable transmission channels, which are still prevalent in India. We have introduced several preprocessing and postprocessing stages to audio and visual processing algorithms to overcome these difficulties. Moreover, the speech and optical character recognition (OCR) technologies for different Indian languages (including Indian English) are under various stages of development under the umbrella of TDIL project [1–5] and are far from a state of maturity. All these factors lead to difficulties in creating a reliable transcript of the spoken or the visual text. We have improved the robustness of the system by restricting the audio-visual processing tasks to discover a small set of keywords of domain interest. These keywords are derived from Really Simple Syndication (RSS) feeds pertaining to the domain of interest. Moreover, these keywords are continuously updated as new feeds arrive and thus, they relate to news stories of contemporary interest. This alleviates the problem of long turn-around time associated with manual updates of the dictionaries, which may fail to keep pace with a fast changing global scenario. We create a multilingual keyword list in English and Indian languages to enable keyword spotting in different TV channels, both in spoken and visual forms. The multilingual keyword list helps us to automatically map the spotted keywords in different Indian languages to their English (or any other language) equivalents for uniform indexing across multiple channels.

The rest of the paper is organized as follows. We review the state-of-the-art in news video analysis in Section 2. Section 3 provides the system overview. Section 4 describes the techniques adopted by us for keyword extraction from speech and visuals from multilingual channels in details. Section 5 provides an experimental evaluation of the system. Finally, Section 6 concludes the paper and provides direction for future work.

2. Related Work

We provide an overview of research in news video analytics in this section to put our work in context. There has been much research interest in automatic interpretation, indexing and retrieval of audio and video data. Semantic analysis of multimedia data is a complex problem and has been attempted with moderate success in closed domains, such as sports, surveillance and news. This section is by no means a comprehensive review on audio and video analytic techniques that has evolved over the past decade, as we concentrate on automated analysis of broadcast video.

Automated analysis, classification and indexing of news video contents have drawn the attention of many researchers in recent times. A video comprising visual and audio components leads to two complementary approaches for automated video analysis. Eickeler and Mueller [6] and Smith et al. [7] propose classification of the scenes into a few content classes based on visual features. A motion feature vector has been computed from the differences in the successive frames and HMM's have been used to characterize the content classes. In contrast, Gauvain et al. [8] proposes an audio-based approach, where the speech in multiple languages has been transcribed and the constituent words and phrases have been used to index the contents of a broadcast stream. Later work attempts to merge the two streams of research and proposes multimodal analysis, which is reviewed later in this section.

A typical news program on a TV channel is characterized by unique jingles at the beginning and the end of the newscast, which provide a convenient means to delimit the newscast from other programs [9]. Moreover, a news program has several advertisement breaks, which need to be removed for efficient news indexing. Several methods have been proposed for TV Commercial (We have used "commercial" and "advertisement" interchangeably in this paper.) detection. One simple approach is to detect the logos of the TV channels [10], which are generally absent during the commercials, but this might not hold good for many contemporary channels. Sadlier et al. [11] describes a method for identifying the ad breaks using "black" frames that generally precedes and succeeds the advertisements. The black frames are identified by analyzing the image intensity of the frames and audio intensity at those time-points. While American and European channels generally use black frames for separation of commercials and programs, it is not so for other geographical regions, including India [12]. Moreover, the heuristics used to ignore the extraneous black frames appearing at arbitrary places within programs are difficult to generalize. Hua et al. [13] have used the distinctive audio-visual properties of the commercials to train an SVM based classifier to classify video shots into commercial and noncommercial categories. The performance of such classifiers can be enhanced with application of the principle of temporal coherence [12]. Six basic visual features and five basic audio features derived context-based features have been used in [13] to classify the shots using SVM and further postprocessing.

The time-points in a streamed video can be indexed with a set of keywords, which provide the semantics of the video-segment around the time-point. Most of the American and European channels accompanied with closed caption text, which are transcripts of the speech, are aligned with the video time-line and provides a convenient mechanism for indexing a video. Where closed captioned text is not available, speech recognition technology needs to be used.

There are two distinct approaches to the problem. In phoneme-based approach [14], the sequence of phonemes constituting the speech is extracted from the audio track and is stored as metadata in sync with the video. During retrieval, a keyword is converted to a phoneme string and this phoneme string is searched for in the video metadata [15]. In contrast, [16] proposes a speaker independent continuous speech recognition engine that can create a transcript of the audio track and align it with the video. In this approach the retrieval is based on the keywords in text domain. The difference is primarily in the way the speech data is transcribed and archived. In the phoneme-based storage, there is no language dictionary used and the speech data is represented by a continuous string of phonemes. While in the later case a pronunciation dictionary is used to convert short phoneme sequences into known dictionary words and the actual phoneme sequence is not retained. Phone level approach is generally more error-prone than word-based approaches because the phoneme recognition accuracies are very poor, typically 40–50%. Moreover, word-based approach provides more robust information retrieval results [17] because in the word-based storage, a speech signal is tagged by at least 3 best (often referred to as *n*-best) phonemes (instead of only one phoneme) at each instance and the word dictionary is used to resolve which sequence of phonemes to use to be able to correlate the speech with a word in the dictionary. Additional sources of information that can be used for news video indexing constitute output from Optical Character Recognition (OCR) on the visual text, face recognizer and speaker identification [18].

Once the advertisement breaks are removed from a news-program, the latter needs to be broken down into individual news stories for further processing. Chua et al. [19] provide a survey of the different methods used based on the experience of TRECVID 2003, which defined news story segmentation as an evaluation task. One of the approaches involve analysis of speech [20, 21], namely, end-of-sentence identification and text tiling technique [22] which involves computing lexical similarity scores across a set of sentence and has been used earlier for story identification in text passages. Purely text-based approach generally yields low accuracy, motivating use of audio-visual features. Identification of anchor shots [23], cue phrases, prosody, and blank frames in different combinations are used together with certain heuristics regarding news production grammar in this approach. A third approach uses machine learning approach where an SVM or a Maximum Entropy classifier classifies a candidate story boundary point based on multimodal data, namely, audio, visual, and text data surrounding the point. While, some of these approaches use a large number of low-level media features, for example, face, motion, and audio classes, some others [24] proposes abstracting low level features to mid-level to accommodate multimodal features without significant increase in dimensionality. In this approach, a shot is preclassified to semantic categories, such as anchor, people, speech, sports, and so forth, which are then combined with a statistical model such as HMM [25]. The classification of shots also helps in segmenting the corpus into subdomains, resulting in more accurate models and hence, improved story-boundary detection. Besacier et al. [26] report use of long pause, shot boundary, audio change (speaker change, speech to music transition, etc.), jingle detection, commercial detection and ASR output for story boundary detection. TRECVID prescribes use of F1 Score [27], the harmonic mean of precision and recall, as a measure of the accuracy. An accuracy of F1 = 0.75 for multimodal story boundary detection has been reported in [22].

Further work on news video analysis extends to conceptual classification of stories. Early work on the subject [23] achieves binary classification shots to a few predefined semantic categories, like "indoors" versus "outdoor", "nature" versus "man-made", and so forth. This was done by extracting the visual features of the key-frames and using a SVM classifier. Higher level inferences could be drawn by observing co-occurrence of some of these semantic levels, for example, occurrence of "sky", "water", "sand", and "people" on a video frame implied a "beach scene". Later work has found that the performance of concept detection is significantly improved by use of multimodal data, namely audio-visual features and ASR transcripts [24]. A generic approach for multimodal concept detection that combines outputs of multiple unimodal classifiers by ensemble fusion has been found to perform better than early fusion approach that aggregates multimodal features into a single classifier. Colace et al. [28] introduced a probabilistic framework for combining multimodal features for classifying the video shots in a few predefined categories using Bayesian Networks. The advantage of Bayesian classifiers over binary classifiers is that the former not only classifies the shots but also ranks the classification. While judicious combination of multimodal improves the performance of concept detection, it has also been observed that use of query-independent weights to combine multiple features performs worst than text alone. Thus, the above approaches for shot classification could not scale beyond a few predefined conceptual categories. This prompts use of external knowledge to select appropriate feature-weights for specific query classes [18]. Harit et al. [29] provide a new approach to use an ontology that can be used to reason with media properties of concepts and to dynamically derive a Bayesian Network for scene classification in a query context. Topic clustering, or clustering news-videos at different times and from different sources is another area of interest. An interesting open question has been the use of audio-visual features in conjunction with text obtained from automatic speech recognition in discovering novel topics [24]. Another interesting research direction is to investigate video topic detection in absence of Automatic Speech Recognition (ASR) data as in the case of "foreign" language news video [24].

3. Framework for Telecast News Analysis

We envisage a system where a large number of TV broadcast channels are to be monitored by a limited number of human monitor. The channels are in English, Hindi (National language of India), and a few other Indian regional

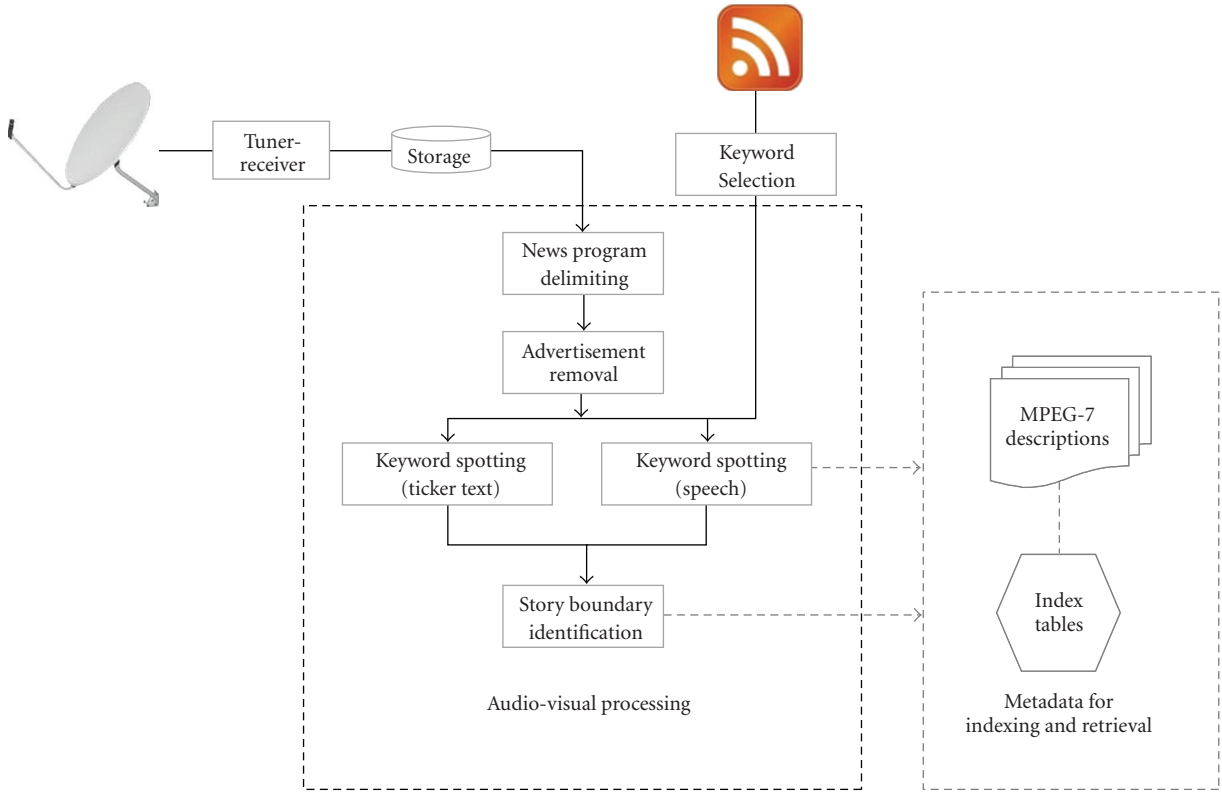


Figure 1: System architecture.

languages. Many of the channels are news channels but some are entertainment channels, which have specific time-slots for news. The contents of the news channels contain weather reports, talk shows, interviews and other such programs besides news. The programs are interspersed with commercial breaks. The present work focuses on indexing news and related programs only.

Figure 1 depicts the system architecture. At the first step of processing, the broadcast streams are captured from Direct to House (DTH) systems and are decoded. They are initially dumped on the disk in chunks of manageable size. These dumps are first preprocessed to identify the news programs. While the time-slots for news on the different channels are known, the accurate boundaries of the programs are identified with the unique jingles that characterize the different programs on a TV channel [9]. The next processing step is to filter out the commercial breaks. Since the black frame-based method does not work for most of the Indian channels, we propose to use a supervised training method [13] for this purpose. At the end of this stage, we get delimited news programs devoid of any commercial breaks.

The semantics of the news contents are generally characterized by a set of keywords (or key phrases) which occur either in the narration of the newscaster or in the ticker text [30] that appears on the screen. The next stage of processing involves indexing the video stream with these extracted keywords. Many American and European channels broadcast transcript of the speech as closed captioned text, which can be used for convenient indexing of the news stream. Since there is no closed captioning available with Indian news channels, we use image and speech processing techniques to detect keywords from both visual and spoken audio track. The video is decomposed into constituent shots, which are then classified into different semantic categories [7, 28], for example, field-shots, news-anchor, interview, and so forth—this classification information is used in the later stages of processing. We create an MPEG-7 compliant content description of the news video in terms of its temporal structure (sequence of shots), their semantic classes and the keywords associated with each shot. An index table of keywords is also created and linked to the content description of the video. The next step in processing is to detect the story boundaries. We propose to use multi-modal cues, visual, audio, ASR output, and OCR data, to identify the story boundaries. We select some of the methods described in [19]. Late fusion method is preferred because of lower dimensionality of features in the supervised training methods and better accuracy [24]. Once the story boundaries are known, analysis of keywords spotted in the story leads to their semantic classification.

In rest of this paper, we deal with the specific problem of indexing the multilingual Indian newscasts with keywords identified in the visuals (ticker text) and in the audio (speech) and improving the indexing performance of news stories with multimodal cues.

4. Keyword-Based Indexing of News Videos

This stage involves indexing of a news video stream with a set of useful keywords and key-phrases (We will use the "keywords" and "key-phrases" interchangeably further in this section.). Since closed captioned text is not available with Indian telecasts, we need to rely on speech processing to extract the keywords. Creating a complete transcript of the speech as in [8] is not possible for Indian language telecasts because of limitations in the speech recognition technology. A pragmatic and more robust alternative is to spot a finite set of contemporary keywords of interest in different Indian languages in the broadcast audio stream. The keywords are extracted from a contemporary RSS feed [31]. We complement this approach with spotting the important keywords in the ticker text that is superimposed on the visuals on a TV channel. While OCR technologies for many Indian languages used for ticker text analysis are also not sufficiently robust, extraction of keywords from both audio and visual channels simultaneously, significantly enhances the robustness of the indexing process.

4.1. Creation of a Keyword File. RSS feeds, made available and maintained by websites of the broadcasting channels or by purely web-based news portals, captures the contemporary news in a semistructured XML format. They contain links to the full-text news stories in English. We select the common and proper nouns in the RSS feed text and the associated stories as the keywords. These proper nouns (typically names of people and places) are identified by a named entity detection module [32] while the common nouns can be identified using frequency count. A significant advantage of obtaining a keyword list from the RSS feeds is the currency of the keywords because of dynamic updates of the RSS feeds. Moreover, the RSS feeds are generally classified into several categories, for example, "business-news" and "international", and it is possible to select the news in one or a few categories that pertains to analyst's domain of interest. Restricting the keyword list to a small number helps in improving the accuracy of the system, especially for keyword spotting in speech.

The English keywords so derived, form a set of concepts, which need to be identified in both speech and visual forms from different Indian language telecasts. While there are some RSS feeds in Hindi and other Indian Languages (For instance, see http://www.voanews.com/bangla/rss.cfm (Bangla), http://feeds.feedburner.com/oneindia-thatstelugu-all (Telugu) and http://feeds.feedburner.com/oneindia-thatshindi-all (Hindi).), aligning the keywords from independent RSS feeds proves to be difficult. We derive the equivalent keywords in Indian languages from the English keywords, each of which is either a proper or a common noun. We use a word level English-to-Indian language dictionary to find the equivalent common noun keywords in an Indian language. We use a pronunciation lexicon (A lexicon is an association of words and their phonetic transcription. It is a special kind of dictionary that maps a word to all the possible phonemic representations of the word.) for transliterating proper names in a semi-automatic matter as suggested in [15]. It is to be noted that (a) the translation of the keyword in English is possible only when the keyword is present in the dictionary else it is transliterated and (b) transliteration of nouns in Indian languages are phonetic and hence there are no transliteration problems that are more visible in a nonphonetic language like English.

```
<RULE NAME="KeyWord">
  <L PROPNAME="keyword">
    <CONCEPT NAME= "Afghanistan">
      <ENG KEY= "Afghanistan">Afghanistan</ENG>
      <BEN KEY= "Afganistan"> আফগানিs ান </ BEN>
      <HIN KEY= "Afganistan">अफगानिस्तान </ HIN>
      <TEL KEY= "Afganistan" >అఫగానిస్తన </ TEL>
    </CONCEPT>
    <CONCEPT NAME= "Rajshekhar">
      <ENG KEY= "Rajshekhar">Rajshekhar</ENG>
      <BEN KEY= "Rajshekhar">বাজশেখব</BEN>
      <HIN KEY= "Rajshekhar">राजशेखर </HIN>
      <TEL KEY= "Rajshekhar"> రాజశేఖర్ </TEL>
    </CONCEPT>
    <CONCEPT NAME= " Terrorist">
      <ENG KEY= "Terrorist ">Terrorist </ENG>
      <BEN KEY= "Santrasbaadi "> সন্ত্রাসবাদী </BEN>
      <HIN KEY= "Atankabaadi ">आतंकबादी </HIN>
      <TEL KEY= "Atankavaadi "> అతన్కవాది </TEL>
    </CONCEPT>
  </L>
</RULE NAME>
```

Figure 2: Keyword list structure.

Finally, the keywords in English and their Indian language equivalents and their pronunciation keys are stored as a multilingual dynamic keyword list structure in XML format. This becomes an active keyword list for the news video channels and is used for both keyword spotting in speech and OCR. We show a few sample entries from a multilingual keyword list file in Figure 2. The first two entries represent proper nouns, the name of a place (Afghanistan) and a person (Rajashekar), respectively. The third entry (terrorist) corresponds to a common noun. In Figure 2 every concept is expressed in three major Indian languages, Bangla, Hindi, and Telugu, besides English. We use ISO 639-3 codes (See http://www.sil.org/iso639-3/.) to represent the languages. KEY entries represent pronunciation keys and are used for keyword spotting in speech. The words in Indian languages are encoded in Unicode (UTF-8) and are used as dictionary entries for correcting OCR mistakes. Each concept is associated with a NAME in English, which is returned when a keyword (speech or ticker text) in any of the languages is spotted either in speech or ticker-text, thus resulting in a built-in machine translation.

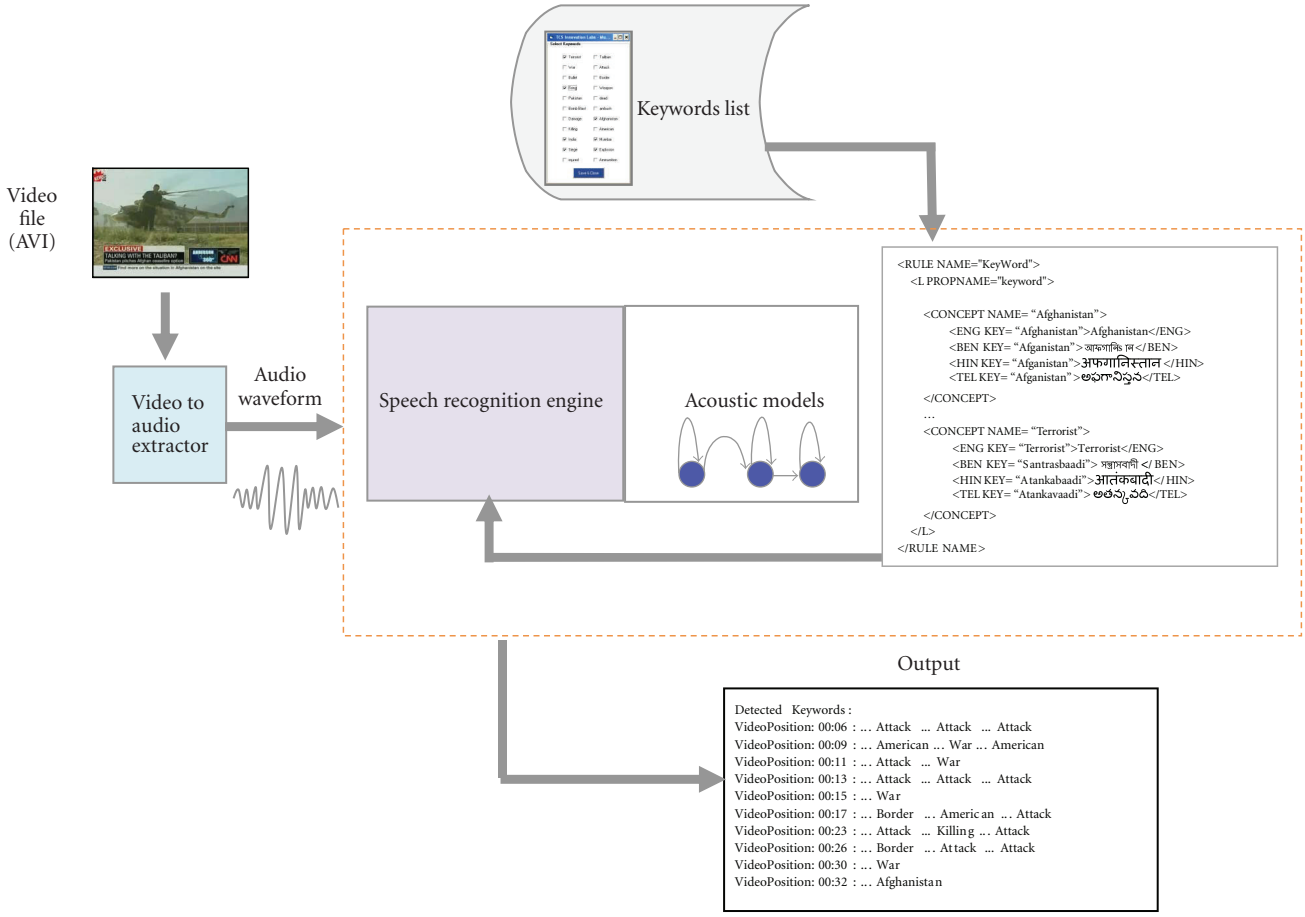


Figure 3: Typical block diagram of a keyword spotting system.

4.2. Keyword Spotting and Extraction from Broadcast News. Audio keyword spotting system essentially enables identification of words or phrases of interest in an audio broadcast or in the audio track of a video broadcast. Almost all the audio keyword spotting systems take the acoustic speech signal (a time sequence, $x(t)$) as input and use a set of (N) keywords or key phrases ($\{K_i\}_{i=1}^{N}$), as reference to spot the occurrences of these keywords in the broadcast [33]. A speech recognition engine ($S : x(t) \rightarrow x(s)$; $x(s)$ is a string sequence $\{s_k\}_{k=1}^{N}$), which is generally speaker independent and large vocabulary, is employed and is ideally supported by the list of keywords that need to be spotted (if $x(s) \in \{K_i\}_{i=1}^{N}$; then S, the speech recognition engine, is deemed to have spotted a keyword). Internally, the speech recognition engine has a built in pronunciation lexicon which is used to associate the words in the keyword list with the recognized phonemic string from the acoustic audio.

A typical functional keyword spotting system is shown in Figure 3. The block diagram shows as a first step the audio track extraction from a video broadcast. The keyword list is the list of keywords or phrases that the system is supposed to identify and locate in the audio stream. Typically this human readable keyword list is converted into a speech grammar file (FSG (finite state grammar) and CFG (context free grammar) are typically grammar used in speech recognition literature.). The speech recognition engine (in Figure 3) makes use of the acoustic models and the speech grammar file to ear mark all possible occurrences of the keywords in the acoustic stream. The output is typically the recognized or spotted words and the time instance at which that particular keyword occurred.

An audio KWS system for broadcast news has been proposed in [34]. The authors suggest the use of utterance verification (using dynamic time warping), out-of-vocabulary rejection, audio classification, and noise reduction to enhance the keyword spotting performance. They experimented on Korean news based on 50 keywords. More recent works include searching multilingual audiovisual documents using the International Phonetic Alphabet (IPA) [35] and transcription of Greek broadcast news using the HMM toolkit (HTK) [36]. We propose a multichannel, multilingual audio KWS system which can be used as a first step in broadcast news clustering.

In a multi channel, multilingual news broadcast scenario the first step towards coarse clustering of broadcast news can be achieved through audio KWS. As mentioned in earlier section broadcast news typically deals with people (including organizations and groups) and places; this makes broadcast news very rich in proper names which have to be spotted in audio. Notice that these words to be spotted

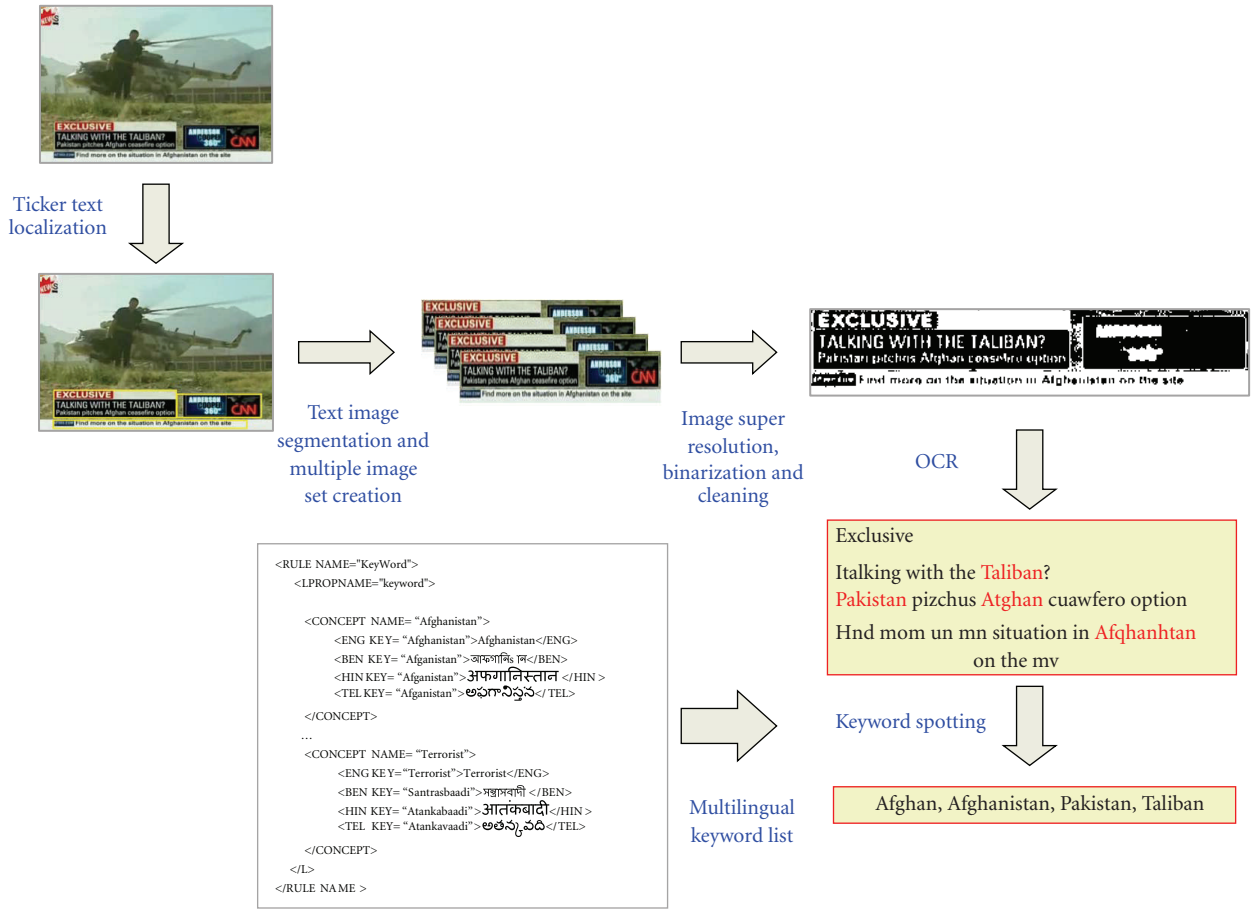


Figure 4: Keyword extraction from ticker text.

are largely language independent, the language independence comes because most of the Indian proper names are pronounced similarly in different Indian languages, implying that the same set of keywords or grammar files can be used irrespective of language of broadcast. In some sense we do not need to (a) identify the language being broadcast and (b) maintain a separate keyword list for different language channels. However, there is a need for a pronunciation dictionary for proper names. Creating a pronunciation lexicon of proper names is time consuming unlike a conventional pronunciation dictionary containing commonly used words. Laxminarayana and Kopparapu [15] have developed a framework that allows a fast method of creating a pronunciation lexicon, specifically for Indian proper names, which are generally phonetic unlike in other languages, by constructing a cost function and identifying a basis set using a cost minimization approach.

4.3. Keyword Extraction from News Ticker Text. News Ticker refers to a small screen space dedicated to presenting headlines or some important news. It usually covers a small area of the total video frame image (approximately 10–15%). Most of the news channels use two-band tickers, each having a special purpose. For instance, the upper band is generally used to display regular text pertaining to the story which is currently on air whereas "Breaking News" or the scrolling ticker on the lower band relates to different stories or displays unimportant local news, business stocks quotes, weather bulletin, and so forth. Knowledge about the production rule of specific TV channel or program is necessary to segregate the different types of ticker texts. We attempt to identify the desired keywords specified in the multilingual keyword list in the upper band, which relates to the current news story in different Indian channels.

Figure 4 depicts an overview of the steps required for keyword spotting in the ticker text. As the first step, we detect the ticker text present in the news video frame. This step is known as text localization. We identify the groups of video frames where ticker text is available and mark the boundaries of the text (highlighted by yellow colored boxes in the figure). The knowledge about the production rules of a channel helps us selecting the ticker text segments relevant to the current news story. In the next step, we extract these image segments from the identified groups of frames. Further, we identify the image segments containing the same text and combine the information in these images to obtain a high-resolution image using image super-resolution technique. We binarize this image and apply touching character segmentation as an image cleaning step.

These techniques help improve the recognition rate of OCR. Finally, the text images are processed by OCR software and desired keywords are identified from the resultant text using the multilingual keyword list. The following subsections give detailed explanation of these steps.

4.3.1. Text Localization in News Video Frames. The text recognition in a video sequence involves detection of the text regions in a frame, recognizing the textual content and tracking the ticker news video in successive frames. Homogeneous color and sharp edges are the key features of texts in an image or video sequence. Peng and Xiao [37] have proposed color-based clustering accompanied with sharp edge features for detection of text regions. Sun et al. [38] propose a text extraction by color clustering and connected component analysis followed by text recognition using a novel stroke verification algorithm to build a binary text line image after removing the noncharacter strokes. A multiscale wavelet-based texture feature followed by SVM classifier is used for text detection in image and video frames [39]. An automatic detection, localization and tracking of text regions in MPEG videos are proposed in [40]. The text detection is based on wavelet transform and modified k-means classifier. Retrieval of sports video databases using SIFT feature-based trademark matching is proposed by [41]. The SIFT based approach is suitable for offline processing in video database but is not a feasible option in real time MPEG video streaming.

The classifier-based approaches have a limitation that if the test data pattern varies from the data used in learning, robustness of the system gets reduced. In the proposed method we have used the hybrid approach where we localize the candidate text regions initially using the compressed domain data processing and process the region of interest in pixel domain to mark the text region. This approach has a benefit over other in two aspects namely robustness and time complexity.

Our proposed methodology is based on the following assumptions.

(1) Text regions have significant contrast with background color.

(2) News ticker text is horizontally aligned.

(3) The components representing texts region has strong vertical edges.

As stated above we have used compressed domain features and time domain features to localize the text regions. The steps involved are as follows.

(1) Computation of Text Regions Using Compressed Domain Features. In order to determine the text regions in the compressed domain, we first compute the horizontal and vertical energies at the sub block (4×4) level and mark the subblocks as text or nontext assuming that the text regions generally possess high vertical and horizontal energies. To mark the high energy regions we first divide the entire video frame into small blocks each of size 4×4 pixels. Next, we apply integer transformation on each of the blocks. We have selected Integer transformation in place of DCT to avoid the problem of rounding off and complexity of floating point operation. We compute the horizontal energy of the subblock by summing the absolute amplitudes of the horizontal harmonics (C_{U0}) and the vertical energy of the subblock by summing the absolute amplitudes of the vertical harmonics (C_{0V}). Then we compute the average horizontal text energy ($E_{\text{Avg\_Hor}}$) and the average vertical text energy ($E_{\text{Avg\_Ver}}$) for each row of subblocks. Lastly we mark candidate rows if both ($E_{\text{Avg\_Hor}}$) and ($E_{\text{Avg\_Ver}}$) exceed threshold value α, where α is calculated as $\mu_E + a\sigma_E$ where "a" is empirically selected by analyzing the mean and standard deviation of energy values observed over a large number of Indian broadcast channels.

(2) Filter Out the Low Contrast Components in Pixel Domain. Human eye is more sensitive in high-contrast regions compared to the low-contrast regions. Therefore, it is reasonable to assume that the ticker-text regions in a video are created with significant contrast with background colour. This assumption is found to be valid in most of the Indian channels. At the next step of processing, we remove all low-contrast components from the candidate text regions identified in the previous step. Finally, the candidate text segments are binarized using Otsu's method [42].

(3) Morphological Closing. The text components sometimes get disjointed depending on the foreground and background contrast and the video quality. Moreover, non textual regions appear as noise in the candidate text regions. A morphological closing operation is applied with rectangular structural elements with dimension of 3×5 to eliminate the noise and indentify continuous text segments.

(4) Confirmation of the Text Regions. Initially we run a connected component analysis for all pixels after morphological closing to split the candidate pixels into n number of connected components. Then we eliminate all the connected components which do not satisfy shape features like size and compactness (Compactness is defined as the number of pixel per unit area.).

Then we compute the mode for x and y coordinates of top left and bottom right coordinates of the remaining components. We compute the threshold as the mode of the difference between the median and the position of all the pixels.

The components, for which the difference of its position and the median of all the positions are less than the threshold, are selected as the candidate texts. We have used Euclidean distance as a distance measure.

(5) Confirmation of the Text Regions Using Temporal Information. At this stage, the text segments have been largely identified. But, some spurious segments are still there. We use heuristics to remove spurious segments. Human vision psychology suggests that eyes cannot detect any event within 1/10th of a second. Understanding of video content requires

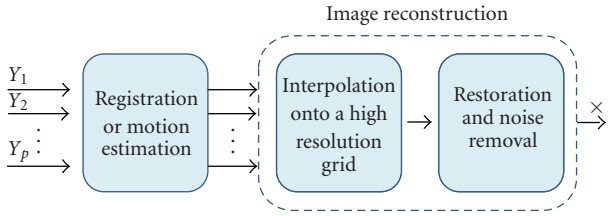


Figure 5: Stages of image super resolution.

at least 1/3rd of a second, that is, 10 frames in a video with frame-rate of 30 FPS. Thus, any information on video meant for human comprehension must persist for this minimum duration. It is also observed that the noise detected as text does not generally persist for significant duration of time. Thus, we eliminate any detected text regions that persists for less than 10 frames. At the end of this phase, we get a set of groups of frames (GoF) containing ticker text. The information together with the coordinates of the bounding boxes for the ticker text are recorded at the end of this stage of processing.

4.3.2. Image Super Resolution and Image Cleaning. The GoF containing ticker text regions cannot be directly used with OCR software because the size of the text is still too small and lacks clarity. Moreover, the characters in the running text are often connected and need to be separated from each other for reliable OCR output.

To accomplish this task we interpolate these images to a higher resolution by using Image Super Resolution (SR) techniques [43, 44] and subsequently perform touching character segmentation as image cleaning process in order to address these problems. The processing steps are given below.

(1) Image Super Resolution (SR). Figure 5 shows different stages of a multiframe image SR system to produce an image with a higher resolution (X) from a set of images $(Y_1, Y_2, \ldots, Y_p)$ with lower resolution. We have used SR technique presented in [45], where information from a set of multiple low resolution images is used to create a higher resolution image. Hence it becomes extremely important to find images with the same ticker text. We perform pixel subtraction of both the images in a single pass. We now count the number of nonblack pixels by using intensity scheme $(R, G, B) < (25, 25, 25)$. We then normalize this count by dividing it by total number of pixels and record this value. If this value exceeds statistically determined threshold "β", we declare the images as nonidentical otherwise we place both the images in the same set. As shown in Figure 5, multiple low resolution images are fed to an image registration module which employs frequency domain approach and estimates the planar motion which is described as function of three parameters: horizontal shift (Δx), vertical shift (Δy), and the planar rotation angle (Φ). In Image Reconstruction stage, the samples of the different low-resolution images are first expressed in the coordinate frame of the reference image. Then, based on these known samples, the image values are interpolated on a regular high-resolution grid. For this purpose bicubic interpolation is used because of its low computational complexity and good results.

Transcription	śivō rakṣatu gīrvāṇabhāṣārasāsvādatatparān
Bengālī	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতৎপরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸ਼ਿਵੋ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oṛiyā	ଶିବଃ। ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ୱାଦତତ୍ପରାନ୍
Tamiḻ	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tĕlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannaḍa	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayāḷam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാൻ
Grantha	ശിവൊ രക്ഷതാ ഗീർവ്വാണഭാഷാരസാസ്വാദതത്പരാൻ

Figure 6: Samples of a few major Indian scripts (Source: http://www.myscribeweb.com/Phrase_sanskrit.png.).

(2) Touching Character Segmentation. We binarize the high-resolution image by Otsu's method [42] containing ticker text. We generally find some of the text characters touching each other in the binarized image because of noise that can adversely affect the performance of the OCR. Hence, we follow up this step with segmentation of touching characters for improved character recognition.

For Touching Character Segmentation, we initially find the average character width for all the characters in the region of interest (ROI) by $\mu_{\mathrm{WC}} = (1/n)\sum_{i=1}^{n} \mathrm{WC}_i$ where "n" is the number of characters in the ROI and "WC_i" is the character width of the ith component. We then compute the threshold for character length and the components with a width greater than that threshold are marked as candidate touching characters. The threshold for character length is computed as $(T_{\mathrm{WC}} = \mu_{\mathrm{WC}} + 3 * \sigma_{\mathrm{WC}})$. We have used $(3 * \sigma_{\mathrm{WC}})$ to ensure higher recall. For our purpose threshold is nearly 64. Then we split them into number of possible touches. The number of touches in a candidate component is computed as the ceiling value of the ratio between actual width and the threshold value, that is, $n_i = [\mathrm{WC}_i/T_{\mathrm{WC}}] + 1$. In some Indian languages (like Bangla and Hindi), the characters in a word are connected by a unique line called *Shirorekha,* also called the "head line". Touching character segmentation for such languages is preceded by the removal of *shirorekha*, which makes character segmentation more efficient.

4.3.3. OCR and Dictionary-Based Correction. The higher quality image obtained as a result of last stage of processing is processed with OCR software to create a transcript of the ticker text in the native language of the channel. The transcript is generally error-prone and we use the multilingual keyword list in conjunction with an approximate string matching algorithm for robust recognition of the desired keywords in the transcript. There are telecasts in English, Hindi (the national language), and several regional languages in India. Many of the languages use their own

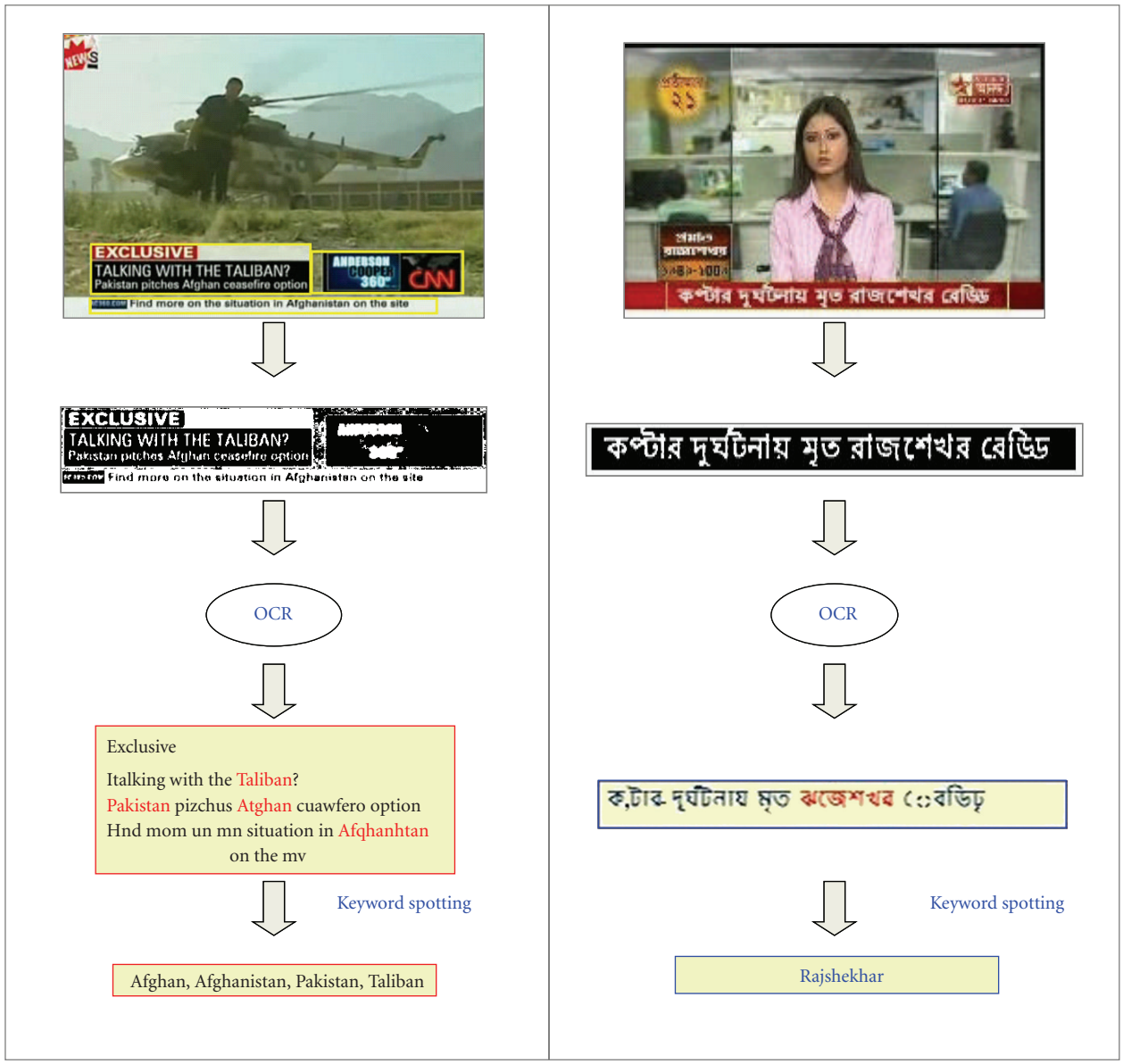


Figure 7: Keyword Identification from English and Bangla news channel.

scripts. Samples of a few major Indian scripts are shown in Figure 6.

The development of OCR in many of these Indian languages is more complex than English and other European languages. Unlike these languages, where the number of characters to be recognized is less than 100, Indian languages have several hundreds of distinct characters. Nonuniformity in spacing of characters and connection of the characters in a word by *Shirorekha* in some of the languages are other issues. There has been significant progress in OCR research in several Indian languages. For example, in Hasnat et al. [46], Lehal [1], and Jawahar et al. [2], word accuracy over 90% has been attained. Still, many of the Indian languages lack a robust OCR and are not amenable to reliable machine processing. For selecting a suitable OCR to work with English and Indian languages, we looked for the highly ranked OCRs identified at The Fourth Annual Test of OCR Accuracy [47] conducted by Information Science Research Institute (ISRI (http://www.isri.unlv.edu/ISRI/)). Tesseract [48] (More information on Tesseract and download packages are available at http://code.google.com/p/tesseract-ocr/.), an open source OCR, finds a special mention because of its reported high-accuracy range (95.31% to 97.53%) for the magazine, newsletter, and business letter test-sets. Besides English, Tesseract can be trained with a customized set of training data and can be used for regional Indian languages. Adaptation of Tesseract for Bangla has been reported in [46]. Thus, we find Tesseract to be a suitable OCR for creating transcripts of English and Indian language ticker text images extracted from the news videos.

Table 1: Results for keyword spotting in speech with master keyword list.

Story id	Instances of keywords present	Keywords found		Retrieval performance		
		True positives	False Positives	Recall (%)	Precision (%)	F-measure (%)
[1]	[2]	[3]	[4]	[5]	[6]	[7]
				[3]/[2] ∗ 100	[3]/([3] + [4]) ∗ 100	2 ∗ [5] ∗ [6]/([5] + [6])
English Channels						
E001	12	2	5	16.67	28.57	21.05
E002	40	10	6	25.00	62.50	35.71
E003	13	2	3	15.38	40.00	22.22
E004	67	8	12	11.94	40.00	18.39
E005	91	6	7	6.59	46.15	11.54
E006	51	7	8	13.73	46.67	21.21
E007	7	1	3	14.29	25.00	18.18
E008	7	1	3	14.29	25.00	18.18
E009	29	10	6	34.48	62.50	44.44
Overall (English)	*317*	*47*	*53*	*14.83*	*47.00*	*22.54*
Bangla Channels						
B001	7	1	0	14.29	100.00	25.00
B002	14	2	5	14.29	28.57	19.05
B003	13	2	1	15.38	66.67	25.00
B004	13	1	7	7.69	12.50	9.52
B005	29	2	7	6.90	22.22	10.53
Overall (Bangla)	*76*	*8*	*20*	*10.53*	*28.57*	*15.38*
Overall	**393**	**55**	**73**	**13.99**	**42.97**	**21.11**

Table 2: Results for keyword spotting in speech with constrained keyword list.

Story id	Instances of keywords present	Keywords found		Retrieval Performance		
		True positives	False Positives	Recall (%)	Precision (%)	F-measure (%)
[1]	[2]	[3]	[4]	[5]	[6]	[7]
				[3]/[2] ∗ 100	[3]/([3] + [4]) ∗ 100	2 ∗ [5] ∗ [6]/([5] + [6])
English Channels						
E001	12	5	4	41.67	55.56	47.62
E002	40	15	3	37.50	83.33	51.72
E003	13	4	1	30.77	80.00	44.44
E004	67	17	6	25.37	73.91	37.78
E005	91	14	8	15.38	63.64	24.78
E006	51	12	5	23.53	70.59	35.29
E007	7	1	0	14.29	100.00	25.00
E008	7	1	0	14.29	100.00	25.00
E009	29	12	4	41.38	75.00	53.33
Overall (English)	*317*	*81*	*31*	*25.55*	*72.32*	*37.76*
Bangla Channels						
B001	7	3	0	42.86	100.00	60.00
B002	14	3	1	21.43	75.00	33.33
B003	13	4	1	30.77	80.00	44.44
B004	13	1	2	7.69	33.33	12.50
B005	29	8	3	27.59	72.73	40.00
Overall (Bangla)	*76*	*19*	*7*	*25.00*	*73.08*	*37.25*
Overall	**393**	**100**	**38**	**25.45**	**72.46**	**37.66**

Table 3: Results for keyword spotting in ticker text with master keyword list.

Story id	No. of distinct ticker texts	Total instances of keywords present	Keywords found			
			On raw frame	On localized text region	After image super-resolution	After dictionary based correction
[1]	[2]	[3]	[4]	[5]	[6]	[7]
English Channels						
E001	5	41	17	19	24	29
E002	4	26	8	9	16	20
E003	4	23	9	10	13	16
E004	6	40	18	19	25	31
E005	4	31	10	13	17	22
E006	7	46	21	23	28	34
E007	4	21	8	9	12	17
E008	1	1	1	1	1	1
E009	5	19	9	9	11	14
Subtotal—English	40	248	101	112	147	184
Retrieval performance—English (%)			*40.73*	*45.16*	*59.27*	*74.19*
Bangla Channels						
B001	3	7	0	0	2	4
B002	3	7	1	1	2	4
B003	5	9	3	3	6	7
B004	3	6	1	1	2	3
B005	5	11	4	4	5	7
Subtotal—Bangla	19	40	9	9	17	25
Retrieval performance—Bangla (%)			*22.5*	*22.5*	*42.5*	*62.5*
Overall retrieval performance (%)			**38.19**	**42.01**	**56.94**	**72.57**

Despite preprocessing of the text images and high accuracy of Tesseract, the output of the OCR phase contains some errors because of poor quality of the original TV transmission. While it is difficult to improve the OCR accuracy, reliable identification of a finite set of keywords is possible with a dictionary-based correction mechanism. We calculate a weighted Levenshtein distance [49] between every word in the transcripts with the words in corresponding language in the multilingual keyword list and recognize the word if the distance is less than a certain threshold "β". The weights in computing the Levenshtein distance is based on visual similarity of the characters in an alphabet, for example, comparison of "l" (small L) and "1" (numeric one) has a lower weight than two other characters, say "a" and "b". We also put a higher weight for the first and the last letters in a word, considering that OCR has a lower error-rate for them because of the spatial separation (on one side) of these characters. Figure 7 shows examples of transcription and keyword identification from news channels in English and Bangla. We map the Bangla keywords to their English (or any other language) equivalents for indexing using the multilingual keyword file.

5. Experimental Results and Illustrative Examples

We have tested the performance of keyword-based indexing with a number of news stories recorded from different Indian channels in English and in Bangla, which is one of the major Indian languages. The news stories chosen pertained to two themes of national controversy, one involving the comments from a popular cricketer and the other involving a visa-related scam. These stories had been recorded over two consecutive dates. Each of the stories is between 20 seconds and 4 minutes in duration. RSS feeds from "Headlines India" (http://www.headlinesindia.com/) on the same dates have been used to create a master keyword-file with 137 English keywords and their Bangla equivalents. In order to test the improvement in accuracy with restricted domain-specific keyword set, we created a keyword file collected from "India news" category, to which the two stories belonged to. This restricted keyword-file contained 16 English keywords and their Bangla equivalents. The restricted keyword set formed was a subset of the master keyword set.

Sections 5.1 and 5.2 present performance of audio and visual keyword extraction, respectively. Section 5.3 present

Table 4: Results for keyword spotting in ticker text with constrained keyword list.

Story id	No. of distinct ticker texts	Total instances of keywords present	Keywords found			
			On raw frame	On localized text region	After image super-resolution	After dictionary-based correction
[1]	[2]	[3]	[4]	[5]	[6]	[7]
English Channels						
E001	5	36	15	17	22	27
E002	4	23	6	7	14	18
E003	4	23	9	10	13	16
E004	6	35	17	19	24	28
E005	4	31	10	13	17	22
E006	7	39	19	21	25	31
E007	4	18	7	8	11	16
E008	1	1	1	1	1	1
E009	5	16	7	7	9	12
Subtotal—English	40	222	91	103	136	171
Retrieval performance—English (%)			*40.99*	*46.40*	*61.26*	*77.03*
Bangla Channels						
B001	3	6	0	0	2	4
B002	3	6	1	1	2	4
B003	5	7	3	3	5	6
B004	3	4	1	1	2	3
B005	5	11	4	4	5	7
Subtotal—Bangla	19	34	9	9	16	24
Retrieval performance—Bangla (%)			*26.47*	*26.47*	*47.06*	*70.59*
Overall retrieval performance (%)			**39.06**	**43.75**	**59.38**	**76.17**

the overall indexing performance on combining audio and visual cues. Section 5.4 presents a few illustrative examples that explain the results.

5.1. Keyword Spotting in Speech. Table 1 presents the results for keyword spotting in speech in the same set of news-stories observed with the master list of keywords. Column [2] represents the number of instances when any of the keywords occurred in the speech. We call keyword spotting to be successful, when a keyword is correctly identified in the time neighborhood (within a +/− 15 ms window) of the actual utterance. Column [3] indicates the number of such keywords for each news story. Column [4] indicates when a keyword is mistakenly identified, though it was actually not uttered at that point of time. We compute the retrieval performances recall, precision and F-measure (Harmonic mean of precision and recall) in columns [5]–[7].

We note that the overall retrieval performance is quite poor, more so for Bangla. It is not surprising because we have used a Microsoft speech engine that is trained for American English. The English channels experimented with were *Indian* channels and the accent of the narrators were quite distinct. We performed the same experiments with the constrained set of keywords. Table 2 presents the results in detail. We note that both recall and precision has significantly improved with the constrained set of keywords, which were primarily proper nouns. The retrieval performance for Bangla is now comparable to that of English. This justifies the use of a dynamically created keyword list for keyword spotting, which is a key contribution in this paper. We note that the precision is quite high (72%), implying that the false positives are low. However, the recall is still pretty low (25%). We will show how we have exploited redundancy to achieve a reliable indexing despite poor recall at this stage.

5.2. Keyword Spotting in Ticker Text. Table 3 depicts a summary of results for ticker text extraction from the English and Bangla Channels tested with master keyword list. Each of the news stories is identified by a unique id in column [1]. Column [2] presents the number of distinct ticker text frames detected in the story. Column [3] indicates the total instances of keywords built from the master keyword list actually present in the ticker text accompanying the story. Columns [4]–[6] show the number of keywords correctly detected when the full-frame, the localized text region and the super-resolution image (of localized text region) are subjected to OCR. Column [7] depicts the number of keywords correctly identified after dictionary-based correction is applied over the OCR result from the super-resolution image of localized text region. We note that the overall accuracy of keyword detection progressively increases from 38.2% to 72.6% through these stages of processing. In Table 3, retrieval performance refers to the recall value. We have observed very few false positives

Table 5: Indexing performance for audio, visual and combined channels.

	Audio			Visual			Combined		
Story id	No. of distinct keywords	Keywords correctly identified	Indexing Performance IP_a (%)	No. of distinct keywords	Keywords correctly identified	Indexing Performance IP_v (%)	No. of distinct keywords	Keywords correctly identified	Indexing Performance IP_o (%)
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
	$\lvert K_a\rvert$	$\lvert K_a\rvert$	[3]/[2] $*$ 100	$\lvert K_v\rvert$	$\lvert k_v\rvert$	[6]/[5] $*$ 100	$\lvert K_o\rvert$	$\lvert K_o\rvert$	[9]/[8] $*$ 100
English channels									
E001	8	5	62.50	13	9	69.23	13	11	84.62
E002	10	7	70.00	9	7	77.78	14	12	85.71
E003	7	5	71.43	10	8	80.00	11	9	81.82
E004	12	9	75.00	13	10	76.92	17	15	88.24
E005	21	12	57.14	11	8	72.73	21	18	85.71
E006	13	9	69.23	15	11	73.33	16	14	87.50
E007	5	2	40.00	10	9	90.00	14	13	92.86
E008	5	2	40.00	1	1	100.00	5	3	60.00
E009	12	9	75.00	9	9	100.00	15	14	93.33
Overall (English)	*93*	*60*	*64.52*	*91*	*72*	*79.12*	*126*	*109*	*86.51*
Bangla channels									
B001	3	2	66.67	4	2	50.00	5	5	100.00
B002	5	4	80.00	4	2	50.00	9	7	77.78
B003	7	4	57.14	6	5	83.33	9	8	88.89
B004	6	3	50.00	3	3	100.00	7	5	71.43
B005	9	6	66.67	5	4	80.00	10	8	80.00
Overall (Bangla)	*30*	*19*	*63.33*	*22*	*16*	*72.73*	*40*	*33*	*82.50*
Overall	**123**	**79**	**64.23**	**113**	**88**	**77.88**	**166**	**143**	**86.14**

(<1%), that is, a keyword mistakenly identified though it is actually not there in the text, and hence we do not present precision in the table. We also observe that the average accuracy for detecting Bangla text with OCR is significantly poor compared to that of the English text, which can be attributed to the OCR performance and quality of visuals, but there is significant improvement after dictionary-based correction.

Similar to audio keyword spotting we performed the same experiments with the constrained set of keywords. Table 4 presents the results in details. We found that by using constrained keywords list the results at every stage have improved, though not as significantly as in the case of speech.

5.3. Improving Indexing Performance by Exploiting Redundancy. While, we have presented the retrieval performance for audio and visual keyword recognition task in the previous sections, the goal of the system is to index the news-stories with appropriate keywords. We define the indexing performance of the system as

$$\text{IP} = \frac{|k|}{|K|} \times 100, \tag{1}$$

where k is the set of distinct keywords correctly identified (and used for indexing the story) and K is the set of distinct keywords present in the story.

The indexing performance is improved by exploiting redundancy in occurrence of keywords in audio-visual forms. In particular, we exploit two forms of redundancy.

(a) The same keywords are uttered several times in a story or appear several times on ticker text. A keyword missed out in one instance is often detected in another instance providing better indexing performance

(b) The same keyword may appear in both audio and visual forms. A keyword often missed in the speech is often detected in visuals and vice-versa. This adds to indexing performance too.

Let K_a and K_v denote the set of distinct keywords actually occurring in the speech and the visuals, respectively, in a news story. Then, $K_o = K_a \cup K_v$ represents the set of keywords appearing in the news-story. Similarly, let k_a and k_v represent the set of distinct keywords detected in the speech and visuals respectively. Then, $k_o = k_a \cup k_v$ represents the set of keywords detected in the news-story. The audio, visual, and overall indexing performance (IP_a, IP_v, and IP_o, resp.)

Stage	Image	OCR output	Keywords spotted
English news story (E004)			
Full frame (binarized)		The comment that star ED the controver Y …galnxn l:nursh¤•1o:Sa•:Mn Tendulkar has man bowled J al uc ua breaking Snlumn Khurmumd: Smzmn 'fuvwdulknr "('|[sAE*g5 Mw; Ima bcwlnd Thuckumy on thm Hamm N; gw :§:::,.·:, ?::~r gx: eé §§§ §¤§¥1s§ ¥	Tendulkar(1)
After text localization		Salman Knumhaco: Sachin Tendulkar has man bowled Bal Tbuckuav	Salman Sachin Tendulkar Bal (4)
After image cleaning and super resolution		Salman Knurshaadz Sachin Tendulkar nas clean howled Bal Thackeray	Salman Sachin Tendulkar Bal Thackeray (5)
After dictionary based correction	----	----	Salman Khursheed Sachin TendulkarBal Thackeray (6)
Bangla news story (B004)			
Fullframe (binarized)		ৈড় য্ \\\ ডুটুষ্ছু ডু-:ডু! জিঞো ঢুঘা ডুডু-- র ।.!ঢুরূ ৈড় ঢু হু- ছু */ ড়ী ৈড়খীয়ঁড়ী 6ঁ ।ো৾ড়-ৈড়. খ ।ঢা।ো।ো।ো।ো।।।ো, ঢ়ানূ -ঁ ৈড়। ঁ ।ঢুঢুড়ি ।ে। ।মুড়ী ।মুঢু৾ড় 44 সিuড়িয় 2 নs র বূ৾ত তৃ।ো৾মূ-ল-সিপিe সঘর্শ ঞী! 'শৃfk ।ঢু খ্।ডু ক্কেলেলেল-ে৾গঞা ঢুছু	None (0)
After text localization		সিu ড়িয় 2 নসুর বূ৾ত তৃ।ো৾মূ-ল-সিপিe ম সংঘর্ন	সিপিe ম (1)
After image cleaning and super resolution		সিu ড়ির 2 নs র- বূ৾ক ডুণসুল-সিপিe ম সংঘর্ন	সিপিe ম ডুণসুল (2)
After dictionary based correction	----	----	সিপিe ম ডুণসুল সংঘর্ষ (3)

Figure 8: OCR outputs at different stages of English and Bangla ticker text processing.

can be measured as

$$\mathrm{IP}_a = \frac{|k_a|}{|k_a|} \times 100, \qquad \mathrm{IP}_v = \frac{|k_v|}{|k_v|} \times 100,$$
$$\mathrm{IP}_o = \frac{|k_o|}{|k_o|} \times 100 \equiv \frac{|k_a \cup k_v|}{|k_a \cup k_v|} \times 100. \tag{2}$$

Table 5 depicts the indexing performance of the audio, the visual and the overall system, with the constrained keyword list. Note that the indexing performances of audio and visual channels, both English and Bangla, are significantly higher than the respective recall values. This is because of the redundancy of occurrence of keywords

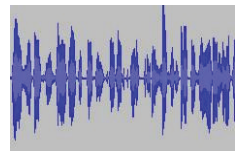
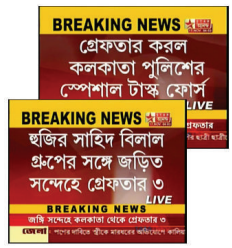
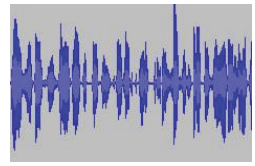
	Keyword spotted in ticker text	Keywords spotted in speech	Combined keyword list
[1]	[2]	[3]	[4]
English news story (E002)			
	Thackeray Sachin Sena Salman sports politics Milkha (7)	Kiran More* Sachin Tendulkar india politics Singh sports (9)	Thackeray Sachin Sena Salman sports politics Milkha Singh Kiran More Tendulkar india (12)
Bangala news story (E005)			
	গিরফ্তার, পুলিস, কলকাতা, হুজি (4)	লs র, গিরফ্তার, কলকাতা, লs র-e-তা বা, বাংলাদেশ, বিতর্ক (6)	গিরফ্তার পুলিস কলকাতা হুজি লs র লs র-e-তা বা বাংলাদেশ বিতর্ক (8)

Figure 9: Combining audio and visual keywords for indexing. *Kiran More: More (pronounced Moré) is a proper noun and not the English word.

in those individual channels. Finally, the overall indexing performance for the stories is greater than the indexing performances of individual audio/visual channels. This is because of the redundancy of keywords across audio and visual channels.

5.4. Illustrative Examples. This section provides some illustrative examples that explain the results in the previous sections. Figure 8 shows the OCR outputs at different stages of processing for examples of English and Bangla ticker text, taken from the stories E004 and B004, respectively. It illustrates the gradual improvement in results through the different stages of image processing and dictionary-based correction.

Figure 9 illustrates improvement in indexing performance by combining audio-visual cues, with an English and a Bangla example. Columns [2] and [3] in the figure show the correctly identified keywords from the ticker text and from speech, respectively. Column [4] depicts the combined keyword list that is used for indexing the story. The combined keyword list is derived as a union of keywords spotted in ticker text and in speech. In these examples, we observe that keywords not detected in speech are often detected in visuals and *vice-versa*. Thus, combining keywords detected in audio and visual forms leads to better indexing performance.

5.5. Comparison. While comparing the system performance, we keep in view the unreliability of the language tools for processing Indian transmission. For example, we have observed the average recall and precision values for keyword spotting in speech to be approximately 15% and 47%, respectively for English (see Table 1), as against typical values of 73% and 85%, respectively in [36]. We also observe that use of a constrained keyword list improves the average recall and precision values to 26% and 72%, respectively (see Table 2), which is still significantly below the reported figures. For keyword detection in ticker text, we have achieved an average recall of 59% (see Table 3) without dictionary-based correction; as compared to 70% reported in [50]. With dictionary-based correction, our recall improves to 67% (see Table 4), which is a reasonable achievement considering complexity of Indian Language alphabets.

An experiment to combine text from speech and visual has been reported in [51]. The authors report recall values for speech recognition and Video OCR as 13% and 6%, respectively. While speech recognition accuracy is comparable to ours, we find the poor OCR results surprising. The authors report a recall of 21% after combining audio and video and dictionary based postprocessing. We have achieved an indexing efficiency of 86%. Though the figures do not directly compare, our system seems to have achieved a much higher performance.

6. Conclusion

We have proposed an architectural framework for automated monitoring of multilingual news video in this paper. The basic idea behind our framework is to combine audio and visual modes to discover the keywords that characterize a particular news-story. Our primary contribution in this paper has been reliable indexing of Indian news telecasts with significant keywords despite inaccuracies of the language tools in processing noisy video channels and deficiencies of language technologies for many Indian Languages. The main contributing factor towards the reliable indexing has been selection of a few domain-specific keywords, in contrast to a complete transcription. Use of several preprocessing and postprocessing stages with the basic language tools has also added to the reliability of results. Moreover, use of RSS feeds to derive the keywords automatically results in contemporariness of the system, which could otherwise be a major operational issue. The conversion of English keywords, which are either proper or common nouns, to their Indian Language equivalents helps indexing non-English transmission with English (or any Indian Language) keywords. The complete end to end solution is made possible by integrating or enhancing available techniques in addition to proposing several techniques that make multilingual, multichannel news broadcast monitoring feasible. The experimental results establish the correctness of the system.

While we have so far experimented with English and one of the Indian languages, namely Bangla, we need to extend the solution to other Indian Languages by integrating appropriate language tools, which are being researched elsewhere in the country. Moreover, India is a large country with twenty-two officially recognized languages and many more "unofficial" languages and dialects. Language tools do not exist and are unlikely to be available in foreseeable future for many of these languages. We propose to direct our future work towards classification of news stories telecast in such languages based on their audio-visual similarity with stories in some reference channels (e.g., some channels in English), which can be indexed using the language technologies.

References

[1] G. S. Lehal, "Optical character recognition of Gurumukhi script using multiple classifiers," in *Proceedings of the International Workshop on Multilingual (OCR '09)*, Barcelona, Spain, July 2009.

[2] C. V. Jawahar, M. N. S. S. K. P. Kumar, and S. S. R. Kiran, "A bilingual OCR for Hindi-Telugu documents and its applications," in *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR '03)*, vol. 1, p. 408, 2003.

[3] E. Hassan, S. Chaudhury, and M. Gopal, "Shape descriptor based document image indexing and symbol recognition," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2009.

[4] U. Bhattacharya and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 444–457, 2009.

[5] S. K. Parui, K. Guin, U. Bhattacharya, and B. B. Chaudhuri, "Online handwritten Bangla character recognition using HMM," in *Proceedings of the International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, 2008.

[6] S. Eickeler and S. Mueller, "Content-based video indexing of TV broadcast news using hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99)*, vol. 6, pp. 2997–3000, March 1999.

[7] J. R. Smith, M. Campbell, M. Naphade, A. Natsev, and J. Tesic, "Learning and classification of semantic concepts in broadcast video," in *Proceedings of the International Conference of Intelligence Analysis*, 2005.

[8] J.-L. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," *Communications of the ACM*, vol. 43, no. 2, pp. 64–70, 2000.

[9] H. Meinedo and J. Neto, "Detection of acoustic patterns in broadcast news using neural networks," *Acustica*, 2004.

[10] C.-M. Kuo, C.-P. Chao, W.-H. Chang, and J.-L. Shen, "Broadcast video logo detection and removing," in *Proceedings of the 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '08)*, pp. 837–840, Harbin, China, August 2008.

[11] D. A. Sadlier, S. Marlow, N. Connor, and N. Murphy, "Automatic TV advertisement detection from MPEG bit stream," *Pattern Recognition*, vol. 35, no. 12, pp. 2719–2726, 2002.

[12] T.-Y. Liu, T. Qin, and H.-J. Zhang, "Time-constraint boost for TV commercials detection," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, vol. 3, pp. 1617–1620, October 2004.

[13] X.-S. Hua, L. Lu, and H.-J. Zhang, "Robust learning-based TV commercial detection," in *Proceedings of the ACM International Conference on Multimedia and Expo (ICME '05)*, pp. 149–152, Amsterdam, The Netherlands, July 2005.

[14] K. Ng and V. W. Zue, "Phonetic recognition for spoken document retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 1, pp. 325–328, 1998.

[15] M. Laxminarayana and S. Kopparapu, "Semi-automatic generation of pronunciation dictionary for proper names: an optimization approach," in *Proceedings of the 6th International Conference on Natural Language Processing (ICON '08)*, pp. 118–126, CDAC, Pune, India, December 2008.

[16] J. Makhoul, F. Kubala, T. Leek, et al., "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338–1352, 2000.

[17] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, no. 1, pp. 5–20, 2000.

[18] T. Chua, S. Y. Neo, K. Li, et al., "TRECVID 2004 search and feature extraction tasks by NUS PRIS," in *NIST TRECVID-2004*, 2004.

[19] T. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *Proceedings of the 12th ACM International Conference on Multimedia (MM '04)*, pp. 656–659, 2004.

[20] A. Rosenberg and J. Hirschberg, "Story segmentation of broadcast news in English, Mandarin and Arabic," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, June 2006.

[21] M. Franz and J.-M. Xu, "Story segmentation of broadcast news in Arabic, Chinese and English using multi-window features," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pp. 703–704, 2007.

[22] M. A. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[23] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchor-person detection for news video parsing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 9, pp. 765–776, 2002.

[24] S.-F. Chang, R. Manmatha, and T.-S. Chua, "Combining text and audio-visual features in video indexing," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 5, pp. 1005–1008, 2005.

[25] L. Chaisorn, T.-S. Chua, and C.-H. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web*, vol. 6, no. 2, pp. 187–208, 2003.

[26] L. Besacier, G. Quénot, S. Ayache, and D. Moraru, "Video story segmentation with multi-modal features: experiments on TRECvid 2003," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '04)*, pp. 221–226, October 2004.

[27] Anonymous, "F1 Score," *Wikipedia—The Free Encyclopedia*, February 2010, http://en.wikipedia.org/wiki/F1_score.

[28] F. Colace, P. Foggia, and G. Percannella, "A probabilistic framework for TV-news stories detection and classification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 1350–1353, July 2005.

[29] G. Harit, S. Chaudhury, and H. Ghosh, "Using multimedia ontology for generating conceptual annotations and hyperlinks in video collections," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '06)*, pp. 211–217, Hong Kong, December 2006.

[30] Anonymous, "News Ticker," Wikipedia—The Free Encyclopedia, February 2010, http://en.wikipedia.org/wiki/News_ticker.

[31] D. Winer, "RSS 2.0 Specification," Wikipedia—The free Encyclopedia, February 2010, http://cyber.law.harvard.edu/rss/rss.html.

[32] S. Kopparapu, A. Srivastava, and P. V. S. Rao, "Minimal parsing key concept based question answering system," *Human Computer Interaction*, vol. 3, 2007.

[33] P. Gelin and C. J. Wellekens, "Keyword spotting for video soundtrack indexing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 299–302, May 1996.

[34] Y. Oh, J.-S. Park, and K.-M. Park, "Keyword spotting in broadcast news," in *Global-Network-Oriented Information Electronics (IGNOIE-COE06)*, pp. 208–213, Sendai, Japan, January 2007.

[35] G. Quenot, T. P. Tan, L. V. Bac, S. Ayache, L. Besacier, and P. Mulhem, "Content-based search in multi-lingual audiovisual documents using the international phonetic alphabet," in *Proceedings of the 7th International Workshop on Content-Based Multimedia Indexing (CBMI '09)*, Chania, Greece, June 2009.

[36] D. Dimitriadis, A. Metallinou, I. Konstantinou, G. Goumas, P. Maragos, and N. Koziris, "GRIDNEWS1a distributed automatic Greek broadcast transcription system," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '09)*, 2009.

[37] J. Yi, Y. Peng, and J. Xiao, "Color-based clustering for text detection and extraction in image," in *Proceedings of the ACM International Multimedia Conference and Exhibition (MM '07)*, pp. 847–850, Augsburg, Germany, Sebtember 2007.

[38] J. Sun, Z. Wang, H. Yu, F. Nishino, Y. Katsuyama, and S. Naoi, "Effective text extraction and recognition for WWW images," in *Proceedings of the ACM Symposium on Document Engineering (DocEng '03)*, pp. 115–117, Grenoble, France, November 2003.

[39] Q. Ye, Q. Huang, W. Gao, and D. Zhao, "Fast and robust text detection in images and video frames," *Image and Vision Computing*, vol. 23, no. 6, pp. 565–576, 2005.

[40] J. Gllavata, R. Ewerth, and B. Freisleben, "Tracking text in MPEG videos," *ACM*, 2004.

[41] A. D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, "Trademark matching and retrieval in sports video databases," in *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR '07)*, pp. 79–86, Augsburg, Germany, Sebtember 2007.

[42] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[43] R. Y. Tsai and T. S. Huang, "Multiple frame image restoration and registration," in *Advances in Computer Vision and Image Processing*, pp. 317–339, JAI Press, Greenwich, Conn, USA, 1984.

[44] V. H. Patil, D. S. Bormane, and H. K. Patil, "Color super resolution image reconstruction," in *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA '07)*, vol. 3, pp. 366–370, 2007.

[45] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–14, 2006.

[46] M. A. Hasnat, M. R. Chowdhury, and M. Khan, "Integrating Bangla script recognition support in Tesseract OCR," in *Proceedings of the Conference on Language and Technology*, 2009.

[47] S. V. Rice, F. R. Jenkins, and T. A. Nartker, "The fourth annual test of OCR accuracy," Tech. Rep. 95-04, Information Science Research Institute, University of Nevada, Las Vegas, Nev, USA, April 1995.

[48] R. Smith, "An overview of the Tesseract OCR engine," in *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR '07)*, vol. 2, pp. 629–633, September 2007.

[49] M. Gilleland, "Levenshtein Distance, in Three Flavors," February 2010, http://www.merriampark.com/ld.htm.

[50] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 4, pp. 256–268, 2002.

[51] A. G. Hauptmann, R. Jin, and T. D. Ng, "Multi-modal information retrieval from broadcast video using OCR and speech recognition," in *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pp. 160–161, Portland, Ore, USA, 2002.