Multimodal indexing of Multi-lingual News Video

Hiranmay Ghosh[1], Sunilkumar Kopparapu[2], Tanushyam Chattopadhyay[3],
Ashish Khare[1], Sujal Wattamwar[1], Amarendra Gorai[1], Megha 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

Abstract

It is important for several agencies to monitor public news telecasts. Monitoring a large number of channels round-the-clock requires huge manpower and is error-prone. The problem is more severe in a country like India, where there are many national and regional language channels, besides English. We present a framework and a set of techniques for automatic identification and indexing of the news stories based on keywords of contemporary interest. These keywords are derived from RSS feed and detected in speech and on ticker text in English and in some Indian languages. Since speech recognition and OCR techniques are not mature in many Indian languages, we use a novel indexing technique for stories on regional language channels based on their audio-visual similarity with other stories. We cluster similar stories and automatically filter out duplicate transmissions of recorded events for optimizing monitoring effort.

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 several agencies, including national security and intelligence services. 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[1] and Nexedia rich media solution[2] 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.

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 and analyze semantics of the broadcast stream. Thus, we need to rely completely on audio-visual processing of the broadcast channels. A major challenge for automated processing of Indian telecast channels is poor signal to noise ratio in audio transmission, and poor resolution (768 x 576) of the video channel. We have used improved audio and video processing algorithms to overcome these difficulties. Another issue with Indian telecasts is absence of black frames, which forces us to explore more sophisticated algorithms for advertisement break and story boundary detection. 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[3] and are far from a state of maturity. Thus, it is not possible to create a reliable transcript of the spoken or the visual text. We have overcome this technology gap by identifying a finite list of keywords and spotting them in speech and ticker-text. Restricting the keywords to a finite vocabulary and spotting them in both audio and visual channels result in robust indexing. These keywords of contemporary interest are derived from Really Simple Syndication (RSS) feeds of news agencies and are dynamically updated. This alleviates the problem of long turn-around time for updating the static and customized dictionaries used in some commercial solutions. 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. Another innovation in our solution is to identify and filter out repeat transmissions based on audio-visual features, which saves valuable analyst time. An interesting aspect of our work is to cluster similar news-stories from multiple channels broadcasting in different languages based on visual features, making the analysis process more efficient. An added benefit of this approach is that the news in “foreign” languages, for which mature OCR and speech technologies do not exist, can also be indexed based on their similarity with stories in some reference 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. Sections 4 and 5 describe some specific technology

[1] http://www.bbn.com/products_and_services/bbn_broadcast_monitoring_system/
[2] http://www.nexidia.com/solutions/rich_media
[3] Technology Development in Indian Languages (TDIL). Ministry of Information Technology, Government of India. http://tdil.mit.gov.in/

elements of the solution, namely (a) keyword extraction from speech and visuals and (b) similar story and repeat transmission detection, where we have concentrated on. Finally, Section 6 concludes the paper.

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 tractable solutions 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 the present times. A video comprise visual and audio components leading to two complementary approaches for automated video analysis. Eickeler and Müller [1] and Smith et al. [2] 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. [3] 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 multi-modal 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 [4]. Moreover, a news program has several advertisement breaks, which need to be removed for efficient news indexing. Significant research on TV Commercial detection and classification has been motivated by its commercial significance as well. Several methods have been proposed for commercial[4] detection. One simple approach is to detect the logos of the TV channels [5], which are generally absent during the commercials, but this might not hold good for many contemporary channels. Sadlier, et al. [6] 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 separation of commercials and programs with black frames holds good for American and European channels, it does not hold good for other regions, including India [7]. Moreover, the heuristics used to ignore the extraneous black frames appearing at arbitrary places within programs are difficult to generalize. [8] use the distinctive audio-visual properties of the commercials to train an SVM based classifier to classify video shots into commercial and non-commercial categories. The performance of such classifiers can be enhanced with application of the principle of temporal coherence [7]. While six basic visual features and five basic audio features has

[4] We have used 'commercial' and 'advertisement' interchangeably in this paper

been used in [8] to classify the shots, [9] have used mid-level features, namely Audio Scene Change Indicator (ASCI) and Image Frames marked with Product Information (IFMPI) as well as existence of black frames and silence, to train the classifier for more accurate determination of the commercial boundaries. Further works on broadcast ad analysis focus on ad classification [9] [10], a detailed review of which is out of scope of the paper.

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 are 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 [11], 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 [12]. In contrast, [13] 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. Phone level approach is generally more error-prone than word based approaches because the phoneme recognition accuracies are very poor. Moreover, word based approach provides more robust information retrieval results [14]. Additional sources of information that can be used for news video indexing constitute output from Video OCR, face recognizer and speaker identification [15].

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. [16] 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 [17, 18], namely, end-of-sentence identification and text tiling technique [19] 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 [20], 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, e.g. face, motion and audio classes, some others [21] proposes abstracting low level features to mid-level to accommodate multimodal features without significant increase in dimensionality. In this approach, a shot is pre-classified to semantic categories, such as anchor, people, speech, sports, etc., which are then combined with a statistical model such as HMM [22]. The classification of shots also helps in segmenting the corpus into sub-domains, resulting in more accurate models and hence, improved story-boundary detection. Besacier, et al. [23] 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 [24], 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 [19].

Further work on news video analysis extends to conceptual classification of stories. Early work on the subject [20] achieves binary classification shots to a few predefined semantic categories, like “indoors” vs. “outdoor”, “nature” vs. “man-made”, etc. 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 [19]. 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. [25] introduce 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 [15]. Harit, et al. [26] 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 [21]. Another interesting research direction is to investigate video topic detection in absence of ASR data as in the case of “foreign” language news video [21].

3. System overview

We envisage a system where a large number of TV broadcast channels are to be monitored by a limited number of human monitored. The channels are in English, Hindi (National language of India) and a few other Indian regional languages. Many of the channels are news channels but some are entertainment channels, which has specific time-slots for news. The contents of the news channels contain weather reports, talk shows, interviews and other such programs besides news. All the programs are generally 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 pre-processed 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 [4]. 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 [8, 9] for this purpose. At the end of this stage, we get delimited news programs devoid of any commercial breaks.

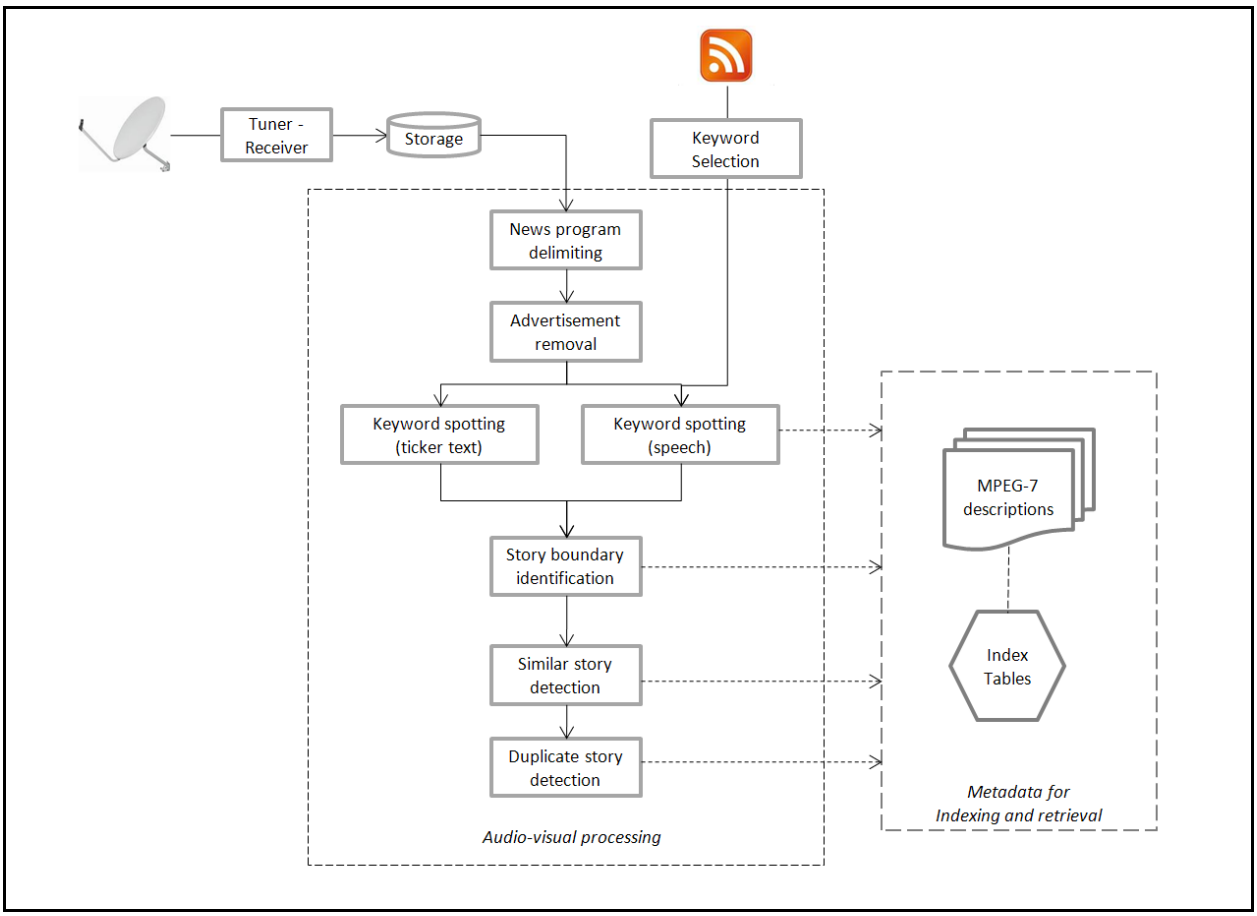

Figure 1: System architecture

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 [27] that appear 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. A possible approach could be to create a complete transcript of the spoken and visual channels for indexing the videos, but the inaccuracy of the ASR and OCR techniques

for many Indian languages do not permit us to do so directly. Instead we derive a small (a few hundred) set of keywords from RSS feeds of some news portals[5]. Use of a limited set of keywords helps in robust detection of keywords in the audio track using speech language grammar models. Another significant advantage of this approach is that we can update the keyword list dynamically based on topics of contemporary interest. The video is decomposed into constituent shots, which are then classified into different semantic categories [2, 25], e.g. filed-shots, news-anchor, interview, etc. – 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 [16]. Late fusion method is preferred because of lower dimensionality of features in the supervised training methods and better accuracy [21]. The final stage of processing involves similar and duplicate story detection. A news story is said to be similar to another if it represents an alternate presentation of the same event, as it generally happens on different channels. Similar stories are characterized by a bulk of visually similar field-shots, though their perspectives (camera angle and distance), durations, sequences and some other properties may differ. In general, similar stories from different channels have different accompanying narrations. A duplicate story (or repeat telecast) is a sub-class of similar stories, where the appearance, duration and sequence of the visuals as well as the narration are identical, though subject to different levels of noise. A comparison of keyword sets discovered in the stories help in identifying candidate similar and duplicate stories and reducing the computational complexity of subsequent audio-visual comparison. Similar stories from different channels are detected based on the visual similarity of the contents. Duplicate stories are identified by stronger check on visual similarity as well as audio similarity on the stories from the same channel on adjacent time-slots. Since OCR and speech processing technologies are not mature for many Indian languages, similar story detection is a powerful tool to index the stories in these “foreign” languages by comparing them with stories on some reference channels which can be indexed by keyword-spotting technique. The content description of the videos and the index tables are further augmented with the information gathered at this stage.

We have used some of the published algorithms to realize the overall system. We concentrate on two topics (a) keyword extraction from speech and visuals and (b) similar story and repeat transmission detection, which required our specific contributions, in the following two sections of this paper.

[5] It could be advantageous if the TV broadcasting channel also hosts a portal with an RSS feed; while this is true of many media houses. In this paper we do not assume this requirement to be mandatory.

4. Keyword Based Indexing of News Videos

This stage involves indexing of news video stream with a set of useful keywords and key-phrases[6]. 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 [3] 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 contemporary RSS feed [28]. 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 mature, 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 [28], made available and maintained by websites of the broadcasting channels or by purely web based news portals, captures the contemporary news in a semi-structured XML format. They contain links to the full-text news stories in English. We apply suitable natural language processing techniques [29] on this text to extract relevant keywords out of it. 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.

The English keywords so derived, form a set of concepts, which need to be identified in speech and in visual forms from Indian language telecasts. The concepts are typically proper and common nouns. We identify their Indian language equivalents by making use of the pronunciation lexicon[7] and a set of dictionaries. In case of common nouns a word level English to Hindi (or a regional language) dictionary is used; on the other hand, we use pronunciation lexicon for proper names. Note that proper nouns require transliteration while common nouns require translation.

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 speech processing 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 names of a place and a person respectively. The third entry corresponds to a common noun. Every concept is expressed in three major Indian languages, Bangla, Hindi and Telugu, besides English. We use ISO 639-3 codes[8] 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,

[6] We shall use the 'keywords' and 'key-phrases' interchangeably further in this section.

[7] 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.

[8] See http://www.sil.org/iso639-3/

which is returned when a keyword (speech or ticker text) in any of the languages are spotted either in speech or ticker-text, thus resulting in an in-built machine translation.

```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

        <CONCEPT NAME="Afghanistan">
             <ENG KEY="Afghanistan">Afghanistan</ENG>
             <BEN KEY="Afganistan">আফগানিস্তান</BEN>
             <HIN KEY="Afganistan">अफगानिस्तान </HIN>
             <TEL KEY="Afganistan">అఫగానిస్తన</TEL>
        </CONCEPT>

        <CONCEPT NAME="Rajshekhar">
             <ENG KEY="Rajshekhar">Rajshekhar</ENG>
             <BEN KEY="Rajshekhar">রাজশেখর</BEN>
             <HIN KEY="Rajshekhar">राजशेखर</HIN>
             <TEL KEY="Rajashekhar">రాజశేఖర్</TEL>
        </CONCEPT>

        <CONCEPT NAME="Terrorist">
             <ENG KEY="Terrorist">Terrorist</ENG>
             <BEN KEY="Santrasbaadi">সন্ত্রাসবাদী </BEN>
             <HIN KEY="Atankabaadi">आतंकबादी</HIN>
             <TEL KEY="Atankavaadi">అతన్కవది</TEL>
        </CONCEPT>
   </L>
</RULE NAME>
```

Figure 2: Keyword List Structure

4.2 Keyword Spotting and Extraction from Broadcast News

Audio keyword spotting system essentially enables identify words or phrases of interest in an audio broadcast or in the audio track of a video broadcast. Almost all the audio keyword spotting (aKWS) systems take the acoustic speech signal (a time sequence, *x(t)*) as input and uses a set of (N) keywords or phrases ($\{K_i\}_{i=1}^{N}$), fed by human, as reference to spot the occurrences of these keywords in the broadcast [30]. A speech recognition engine (S: x(t) → 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 or phrase that need to be spotted (if x(s) $\in \{K_i\}_{i=1}^{N}$; then S 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[9]. The speech recognition engine (in Figure 3, it is the Microsoft speech recognition engine) 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 aKWS system for broadcast news has been proposed in [31]. 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 multi-lingual audiovisual documents using the International Phonetic Alphabet (IPA) [32] and transcription of Greek broadcast news using the HMM toolkit (HTK) [33]. We propose a multi-channel, multi-lingual aKWS 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 aKWS. Observe that broadcast news typically deals with people (including organizations and groups) and places; this makes broadcast news very rich in proper names. This observation has two implications

(1) The word to be spotted in multilingual channels is essentially language independent (an advantage)
(2) There is a need for a strong pronunciation dictionary for proper names (a disadvantage)

The advantage of this observation is 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, creating a pronunciation lexicon of proper names is time consuming unlike a conventional language pronunciation dictionary. Laxminarayana, et al. [34] have developed a framework that allows a fast method of creating a pronunciation lexicon, specifically for proper names, by constructing a cost function and identifying a basis set using a cost minimization approach.

[9] FSG (finite state grammar) and CFG (context free grammar) are typically grammar used in speech recognition literature

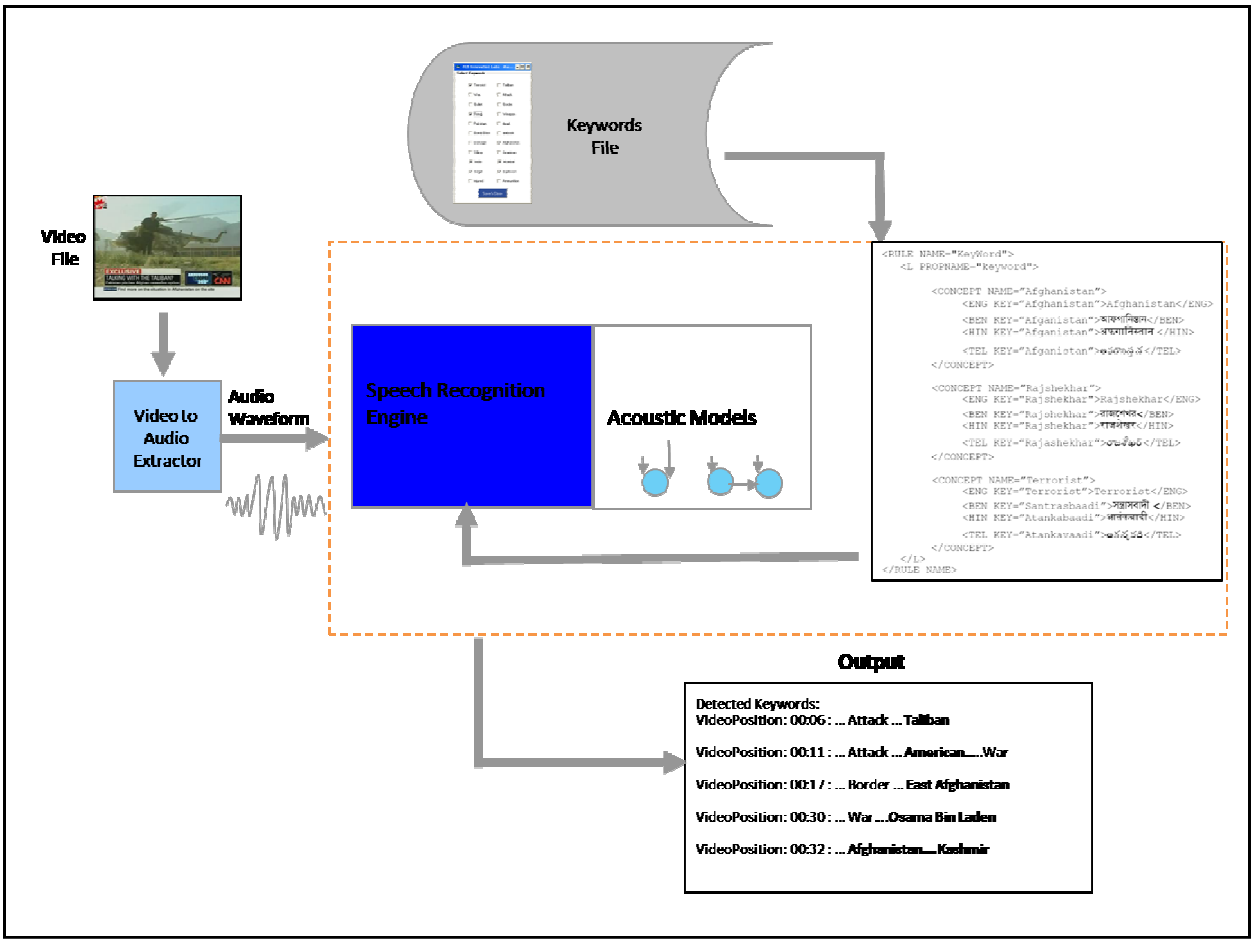

Figure 3: Typical block diagram of a keyword spotting system

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, etc. 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 ticket 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 using a dynamic threshold 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 sub-sections give detailed explanation of these steps.

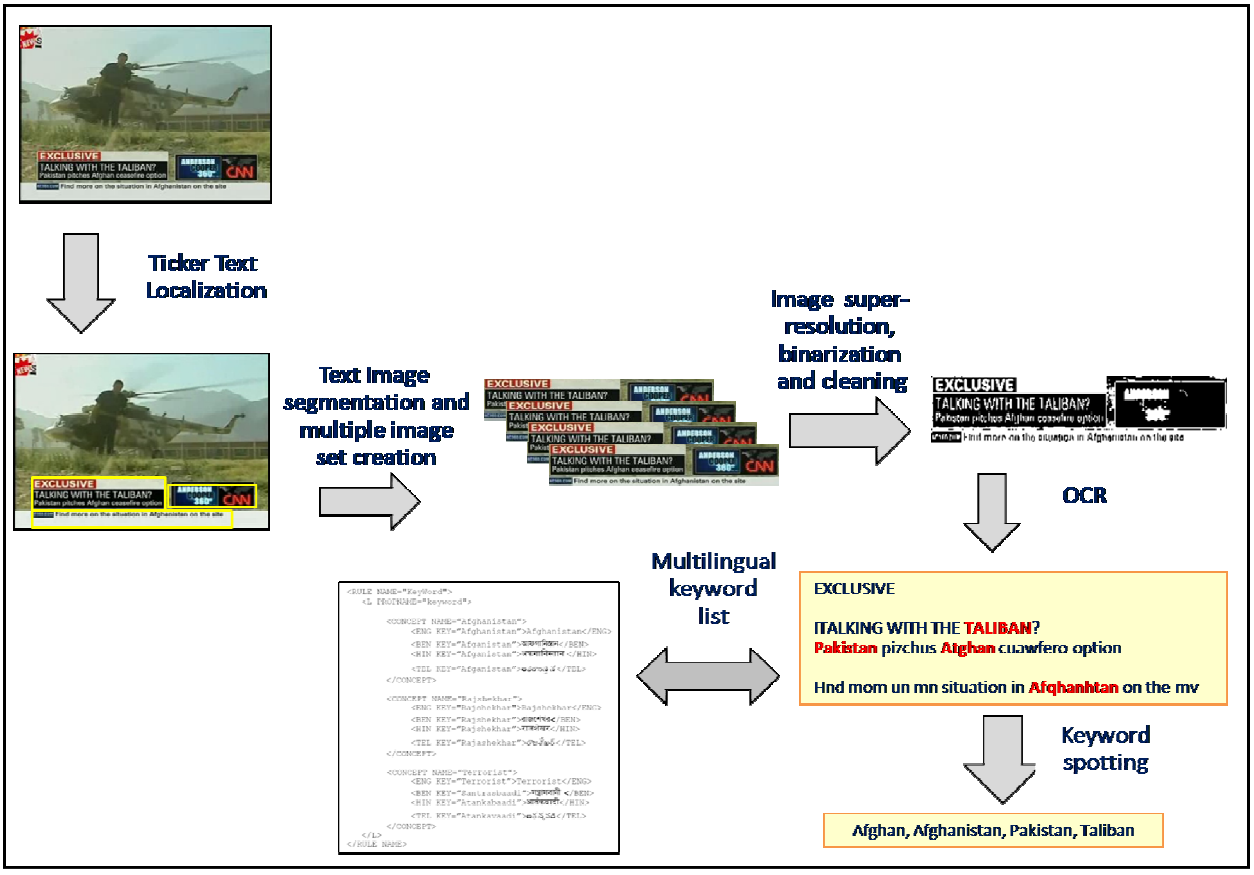

Figure 4: Keyword Extraction from Ticker-Text

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 [35] have proposed color based clustering accompanied with sharp edge features for detection of text regions. Sun, et al. [36] proposes 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 non-character strokes. A multi-scale wavelet based texture feature followed by SVM

classifier is used for text detection in image and video frames [37]. An automatic detection, localization and tracking of text regions in MPEG videos are proposed in [38]. 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 [39]. 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 has 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 compressed domain, we first compute the horizontal and vertical energies at the sub block (4x4) level and mark the sub-blocks as text or non text 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 4x4 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 floating point operation. Then we compute the horizontal energy (E_{hor}) of the sub-block by summing the absolute amplitudes of the vertical harmonics and the vertical energy (E_{ver}) of the sub-block by summing the absolute amplitudes of the vertical harmonics.

Finally, we compute the average vertical text energy and the average horizontal text energy for each row of sub-blocks. Average energies of a typical video frame are shown in Figure 5.

Lastly we mark candidate rows for which the average vertical text energy and the average horizontal text energy exceed some threshold value based on statistical analysis. Candidate text regions satisfying the above conditions for the same frame are shown in Figure 6.

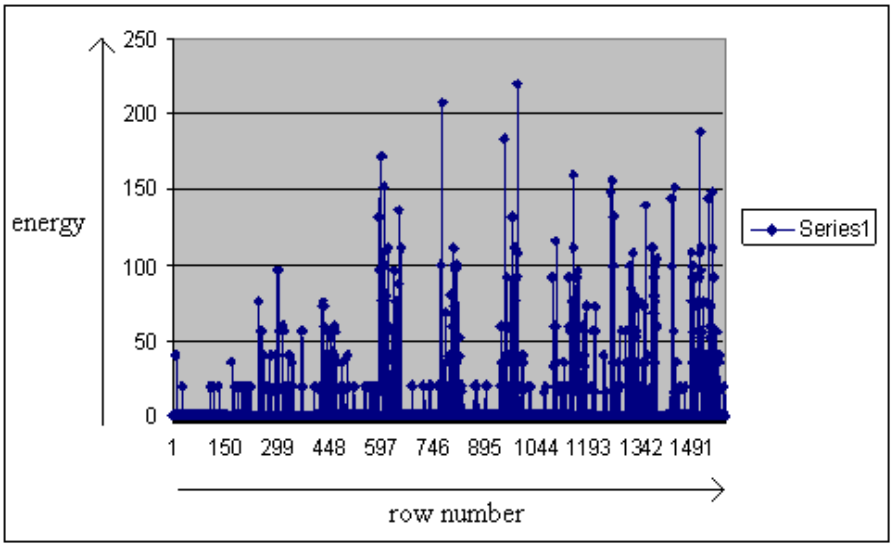

Figure 5: Graph showing energy of the rows

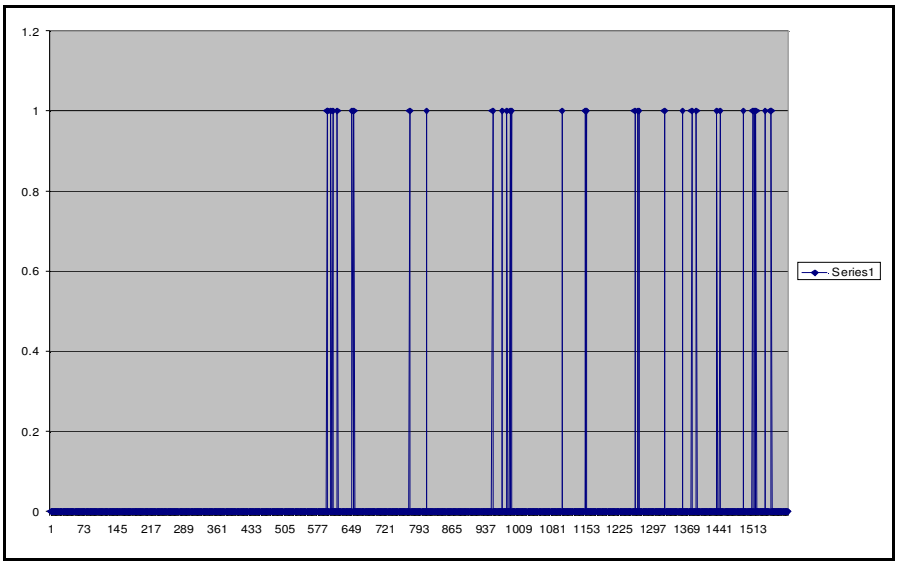

Figure 6: Graph showing higher energy regions as text after applying threshold

(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.

We compute the absolute difference of intensity (I_{diff}) of the neighbouring pixels and then mark all

pixels for which $I_{diff} > \text{THRESHOLD\_CONTRAST}$ to get the binarized output video, where $\text{THRESHOLD\_CONTRAST}$ is a statistically obtained value.

(3) Morphological closing

Because of the poor video quality in Indian telecast, the text components generally get disjointed. 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 3x5 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[10].

Then we compute the mode for *x* and *y* coordinate 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 is less than the threshold, are selected as the candidate texts. We have used Euclidian 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 at least 1/3rd of a second, i.e. 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 [40, 41] and subsequently perform touching character segmentation as image cleaning process in order to address these problems. The processing steps are given below:

[10] Compactness is defined as the number of pixel per unit area.

(1) Image Super Resolution (SR)

Figure 7 shows different stages of a multi-frame image SR system to produce an image with a higher resolution (X) from a set of images (Y1, Y2…Yp) with lower resolution. We have used SR technique presented in [42], 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 same ticker text. We perform pixel subtraction of both the images in a single pass. We now count the number of non-black 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 non identical otherwise we place both the images in same set. As shown in Figure 7, 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.

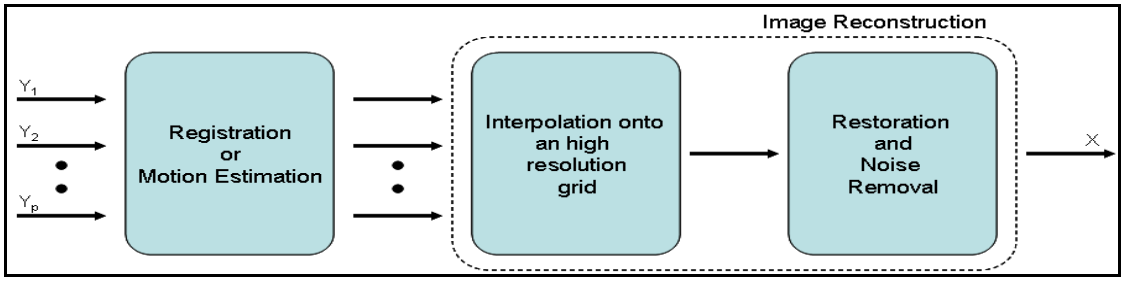

Figure 7: Stages of Image Super Resolution

(2) Touching Character Segmentation

We binarize the high resolution image containing ticker text. We use an adaptive thresholding algorithm [43] to get best possible clarity of the 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). Then we find the mean and average of width of each component. Using the statistical operation we compute the threshold for character length and the components with a width greater than that threshold are marked as candidate touching characters. 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. 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.

We get a higher resolution binarized image with well separated characters, which will now be given to OCR as input.

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 scripts. Samples of a few major Indian scripts are shown in Figure 8.

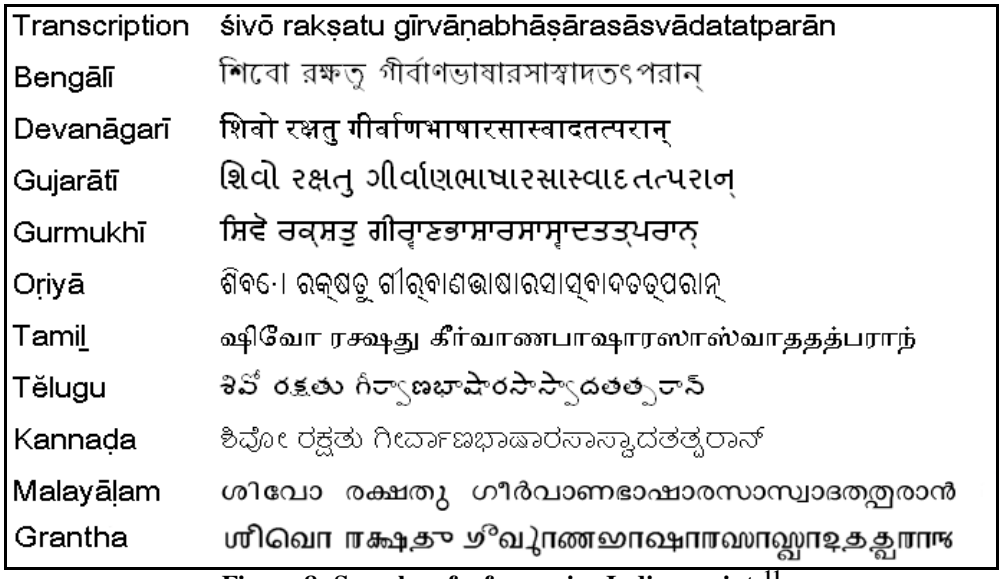

Figure 8: Samples of a few major Indian scripts[11]

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. Non-uniformity in spacing of characters and connection of the characters in a word by *Shirorekha* (head line) in some of the languages are other issues. There has been significant progress in OCR research in several Indian languages, e.g. Bangla [44], Punjabi [45], Hindi and Telugu [46] and word accuracy over 90% has been attained in them. 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[12]). Tesseract [48][13], an open source OCR, finds a special

[11] Source: http://www.myscribeweb.com/Phrase_sanskrit.png
[12] http://www.isri.unlv.edu/ISRI/
[13] More information on Tesseract and download packages are available at http://code.google.com/p/tesseract-ocr/

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 claims to support regional Indian languages also, e.g. Bangla [44], and can be trained with a customized set of training data. 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.

Despite pre-processing 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 9 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.

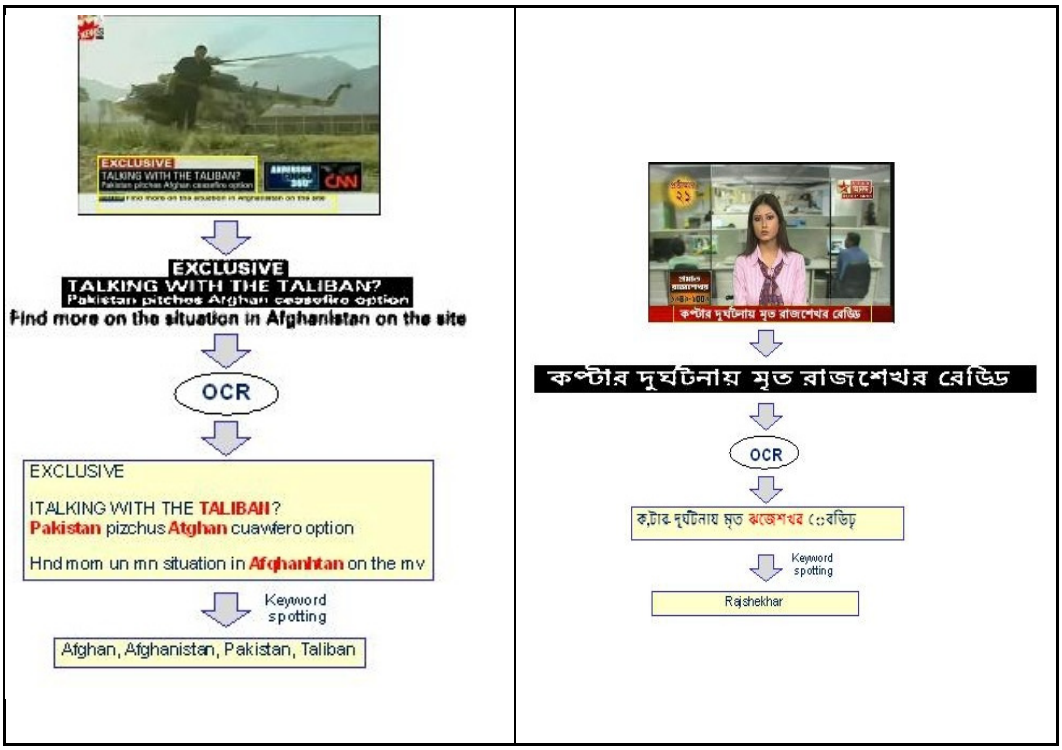

Figure 9: Keyword Identification from English and Bangla news channel

5. Similar and duplicate story detection

This is the last stage of processing and several operations have already been performed on the telecast streams in previous stages. At this stage, we assume that the story boundaries in the news programs have been identified and each story is described with its constituent shots, which are characterized by the media-times and representative frames. Each story is also indexed with a set of English keywords following the process described in Section 4. We use audio-visual analysis to identify similar and duplicate stories. Detection of similar stories and repeat telecasts based on audio-visual analysis is extremely compute intensive and need to be restricted to a smaller subset of stories. Similar stories are likely to be telecast on different channels within few hours of each other and on the same channel on successive news bulletins. Thus, we restrict the candidate stories for audio-visual analysis to this subset. Further, similar stories are likely to be indexed using similar keywords. We characterize each story with a set of keywords and perform a vector-space similarity evaluation. We restrict the candidate set further to the stories, where the similarity score is above a certain threshold.

5.1 Similar story detection

Similar stories on different channels are characterized by similar visuals of field reports, but unique narration of the respective news reader. Thus, similar story detection needs to be based on visual comparison only. The field shots depicting the event are generally sourced from different cameras, often in close vicinity of each other, implying some difference in perspective. These field shots are generally shown in different sequences and for different durations on different channels. Near-duplicate video detection algorithms [50] assume same sequence and duration of the visuals, albeit change of perspective and cannot be used in this context.

To detect the similarity between two stories through visual comparison, the shots comprising the stories are clustered based on the visual similarity of their representative frames. Each cluster in a story is now compared with every cluster in the other story by comparing the central representative frames in the clusters using a visual comparator.

Let $\{c_{11}, c_{12} \ldots c_{1m}\}$ be the clusters in story s1, and $\{c_{21}, c_{22} \ldots c_{2n}\}$ be the clusters in story s2. Let k_{ij} be the number of shots in j^{th} cluster of story i. The process is repeated with every pair of candidate similar stories and clusters of similar stories are discovered. Let $\Gamma_{ij}(c_{1i}, c_{2j})$ represents the match between cluster pair c_{1i} and c_{2j}. Γ_{ij} =1 if the shots in the pair are similar, and 0 otherwise.

We define similarity between the two videos as $sim_{12} = \sum_{ij} (k_{1i}+k_{2j}). \Gamma_{ij} / \sum_{ij} k_{ij}$

The measure is the ratio of number of matching shots and the total number of shots in the two stories. If sim_{12} is greater than a certain threshold (τ), we designate the stories to be similar. As the stories have already been identified as candidates for similarity based on keywords, the threshold τ can be kept quite small. With empirical studies, we set $\tau = 0.4$.

We choose PCA-SIFT feature [51] for visual comparison because of its tolerance towards changes in scale and perspective for image matching. It may also be noted that, if there are significant number of

matching studio shots (portraying a similar studio layout), the method may result in false positives. Restricting the candidate stories with keyword matches alleviates this possibility.

Finally, we enrich the indexing keyword set of each of the stories with the keywords of the other similar stories. Let S= $\{s_1, s_2 \ldots s_n\}$ be a set of similar stories and let $\{K_1, K_2, \ldots K_n\}$ be their respective set of indexing keywords. We associate a set of keywords $K = \{K_1 \cup K_2 \cup \ldots \cup K_n\}$ to every story $s_i \in S$.

We present a couple of illustrative examples for similar story detection.

(1) Similar story occurring in different news channels in same language

Figure 10 and Figure 11 shows two candidate similar stories, broadcasted by two different TV channels. Both of these stories are related to *26/11 Terrorist attack in Taj Hotel Mumbai, India.* Figures 10 and 11 depict representative frames for the shots in these stories.

Figure 10: (Broadcaster: CNN IBN, language: English, number of shots: 6, duration: 2:28 min)

Figure 11: (Broadcaster: NDTV, language: English, number of shots: 9, duration: 3:31 min)

Figure 12 shows the clusters of shots in the two stories. The similarity between clusters C1 and C2 from story 1 and clusters C2 and C1 from story 2 respectively is shown by arrow marks in the figure.

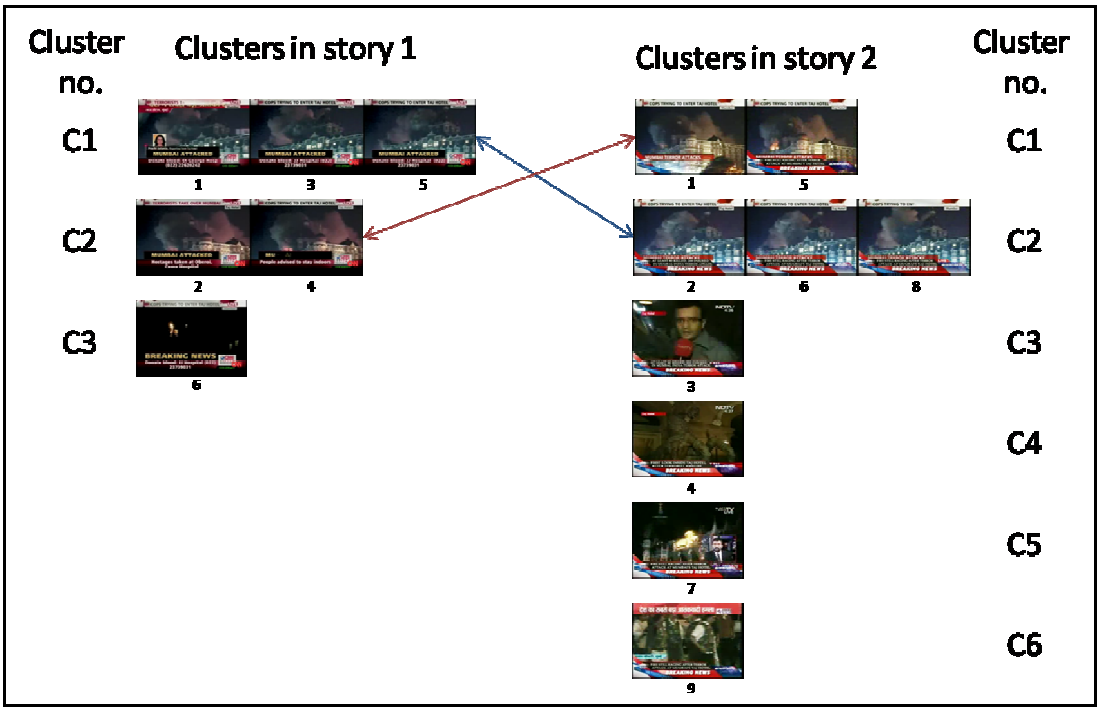

Figure 12: Clustering of shots in stories (same language) and matching between them

As reasonable numbers of shots from story 1 are matched with those of story 2, the two stories are declared to be similar.

(2) Similar stories in different language channels

Figure 13 and Figure 14 shows two similar news stories in two different languages: Telugu and Bangla respectively.

Figure 13: (Broadcaster: TV9, language: Telugu, number of shots: 8, duration: 3:07 min)

Figure 14: (Broadcaster: Star Anand, language: Bangla, number of shots: 5, duration: 2:53 min)

Figure 15 shows the clusters of shots in the two stories shown in Figure 13 and Figure 14. The similarity between cluster C2 from story 1 and cluster C3 story 2 is shown by an arrow mark in the Figure 15.

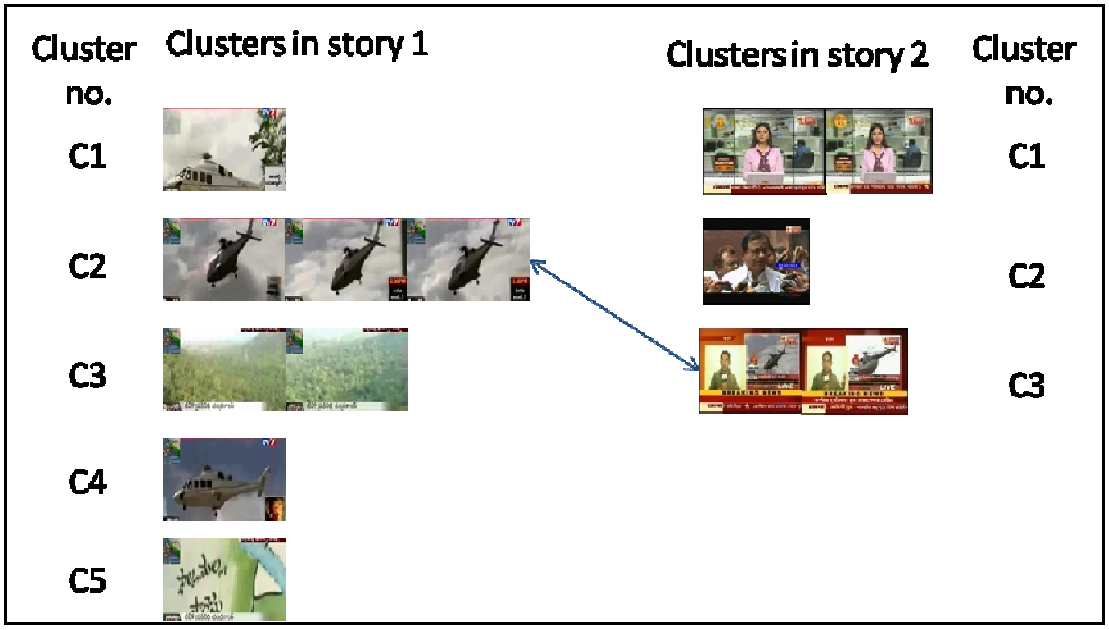

Figure 15: Clustering of shots in two stories (different languages) and matching between them

Presenting the similar stories from a plethora of multilingual channels together in a cluster helps optimizing monitoring efforts in a large way. Moreover, the similar stories share each other's keywords to achieve enhanced indexing. In particular, stories in those languages, for which language tools are not adequately mature benefit significantly from this approach.

5.2 Duplicate story detection

A news story is often repeated several times on a channel. It is desirable to filter the duplicate transmissions to save the efforts of the monitoring agency. Duplicate stories are a subset of similar stories, with tighter matching criteria. Two stories are said to be duplicates of each other only if they are of equal duration and their visual as well as audio representations match exactly.

Let s_1 and s_2 be two news stories. Let T_1 and T_2 be their total durations, and *m* and *n* be the total number of shots in the two stories respectively. Let {h11, h12 … h1m} and {h21, h22, … h2n} represent the shots in the stories s_1 and s_2 respectively and $\{t_{11}, t_{12}, \ldots t_{1m}\}$ and $\{t_{21}, t_{22}, \ldots t_{2n}\}$ be their respective durations.

We call the stories s_1 and s_2 to be visually duplicates if and only if all of the following conditions are satisfied.

(1) $m = n$
(2) $\forall i=1..m,\ t_{1i} = t_{2i}$
(3) $\forall i=1..m,\ h_{1i} \in c_{1p},\ h_{2i} \in c_{2q} \rightarrow \Gamma(c_{1p}, c_{2q}) = 1$

These condition means that there are equal number of shots in the two stories and that each shot in s_1 is of equal duration and visually similar to the corresponding shot in s_2. The different shots of the stories have been compared and clustered in the previous stage of detecting similar stories and we reuse the results. Thus, this step does not involve any visual comparison overheads.

The audio patterns of two stories are matched using the audio fingerprinting method [52]. The audio fingerprints based on perceptual features of audio that are invariant, at least to certain degree, with respect to signal degradations. The fingerprints are matched for each frame block, which is a group of frames, from two streams. The two streams are duplicates if the fingerprints of all the frame blocks are matched. As the streams can be from different channels, they may not match exactly at the desired points. The match may occur at few samples before or after the desired point. Thus the audio frames are matched in a window of some pre-determined size. The window contains couple of audio samples before and after the desired point.

Figure 16 shows audio streams of two different news programs. The information about story boundaries is obtained from the MPEG-7 file. The audio streams of *story2* of program 1 and *story 1* of program 2 are found to be duplicates of each other.

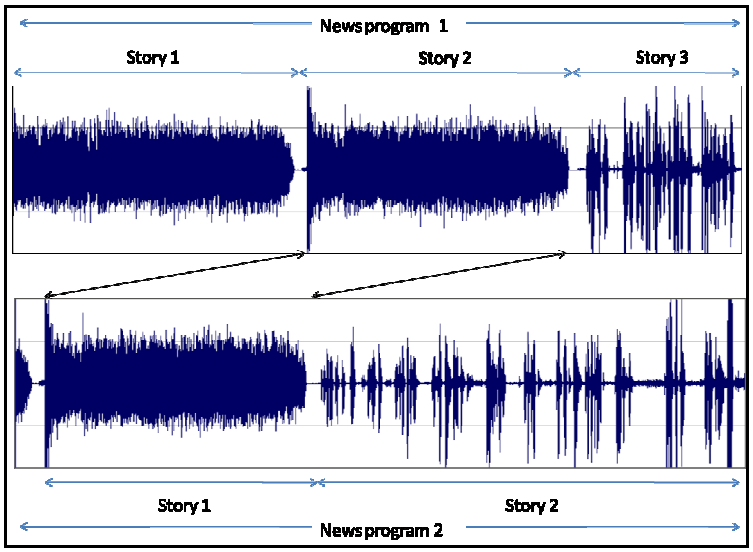

Figure 16: Duplicate stories in audio domain

Removal of the repeat telecasts help largely in optimizing efforts in monitoring telecasts.

6. Conclusion

News broadcast monitoring is a major activity undertaken by most governments to safeguard themselves from external and internal security threats. Continuous monitoring of all the broadcast channels in different languages by human is both time consuming and subject to fatigue and hence erroneous. We proposed a complete end-to-end platform which makes 24x7 monitoring of multi-language, multi-channel TV news broadcast not only manageable but also feasible. The feasibility arises because of synergetic use of not only multiple (audio, video, text) sources of information that exist in the news broadcast but also the use of orthogonal cues (RSS feed) to annotate news broadcast robustly. The main contributions of the paper are in terms of (a) use of visual cues to group together news footage in different languages; (b) use of processed RSS feed in English from multiple sources ; (c) use of RSS feed to extract keywords and key phrases; (d) generation of a set of dynamic keywords in multiple languages using a language dictionary and/or a speech lexicon dictionary; (e) use of a common keyword file to simultaneously assist in enhancing the performance of spotting keywords in both the audio track and OCR of ticket text; and (f) use of audio (proper names) keywords to cluster similar news broadcast followed by visual similarity to filter out similar story broadcast. The complete end to end solution is made possible by integrating or enhancing available technique in addition to proposing several techniques that make multi-lingual, multi-channel news broadcast monitoring feasible. Though we have presented examples from Indian scenario, the proposed solution can be used in any multi-language, multi-channel news broadcast monitoring.

References

1. S. Eickeler and S. Muller, "Content-based video indexing of TV broadcast news using hidden Markov models," In *Proc. IEEE Int Conf on Accoustics, Speech and Signal Processing, 15-19 (ICASSP'99)* Vol 6, pp 2997—3000, March, 1999.

2. J. R. Smith, M. Campbell, M. Naphade, A. Natsev and J. Tesic, "Learning and classification of semantic concepts in broadcast video," *International conference of Intelligence Analysis,* 2005.

3. J. Gauvain, L. Lamel and G. Adda, "Transcribing broadcast news for audio and video indexing", *Communications of the ACM (CACM)*, 43(2), pp 64—70, February 2000.

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

5. C. Kuo, C. Chao, W. Chang and J. Shen, "Broadcast Video Logo Detection and Removing," *International Conference on Intelligent Information Hiding and Multimedia Signal Processing, (IIHMSP '08)* Harbin, 15-17, pp: 837—840, August, 2008.

6. D. A. Sadlier, S. Marlow, N. Connor and N. Murphy, "Automatic TV advertisement detection from MPEG bitstream," *Pattern Recognition*, 35(12), pp 2719 — 2726, December 2002.

7. T. Liu, T. Qin and H. Zhang, "Time-constraint boost for TV commercial detection," *International Conference on Image Processing, (ICIP '04)*, Vol 3, pp: 1617 – 1620, 24-27 October 2004.

8. X. Hua, L. Lu and H. Zhang, "Robust learning-based TV commercial detection," *Proc. 14th ACM International Conference on Multimedia and Expo (ICME)*, Amsterdam, 6 July 2005.

9. L. Duan, J. Wang, Y. Zheng, H. Lu and J. S. Jin, "Segmentation Categorization and identification of commercials from TV streams using multimodal analysis," *International Multimedia Conference (MM'06)*, 23-27 October, 2006.

10. J. Wang, L. Duan, L. Xu, Y. Zheng, J. S. Jin, H. Lu and C. Xu, "TV ad video categorization with probabilistic latent concept learning," *Multimedia Information Retrieval (MIR'07)*, Augsburg, pp 217—226, Sept 2007.

11. N.G. Kennedy, and V. W. Zue, "Phonetic recognition for spoken document retrieval," *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 325–328, 1998.

12. M. Laxminarayana and S. Kopparapu, "Semi-Automatic Generation of Pronunciation Dictionary for Proper Names: An Optimization Approach", *Proceedings of ICON-2008: 6th International Conference on Natural Language Processing*, CDAC, Pune, India, pp. 118-126, Dec 20-22 2008.

13. J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz and A. Srivastava, "Speech and Language Technologies for Audio Indexing and Retrieval," *Proceedings of the IEEE, 88(8)*, August 2000.

14. S. Renals, D. Abberley, D. Kirby and T. Robinson. "Indexing and Retrieval of Broadcast News", *Speech Communication*, vol. 32, pp. 5—20, 2000.

15. T. Chua, S.Y. Neo, K. Li, G.H. Wang, R. Shi. M Zhao, H. Xu S. Gao and T.L. New, "*TRECVID 2004 search and feature extraction tasks by NUS PRIS,*" *In NIST TRECVID-2004*, November 2004.

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

17. A. Rosenberg and J. Hirschberg, "Story segmentation of broadcast news in English, Mandarin and Arabic," *Proc. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 4-9 June 2006.

18. M. Franz and J. Xu, "Story segmentation of broadcast news in Arabic, Chinese and English using multi-window features," *Proc 30th annual international ACM SIGIR conference on research and development in information retrieval (poster),* pp 703 – 704, 2007.

19. M.A. Hearst, "Texttiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, 23(1), pp. 33—64, 1997.

20. X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchor-person detection for news video parsing," *IEEE Trans. Circuits and Systems for Video Technology,* 12(9), pp. 765 – 776, 2002.

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

22. L. Chaisorn, T. Chua and C. Lee, "A multi-modal approach to story segmentation for news video," *World Wide Web: Internet and Web Information Systems,* Vol 6, pp 187—208, 2003.

23. L. Besacier, G. Quénot, S. Ayache and D. Moraru, "Video story segmentation with multi-modal features: experiments on TRECVID 2003", *Multimedia Information Retrieval (MIR'04)*, October 15-16, 2004.

24. Anonymous, "F1 Score", *Wikipedia – The free Encyclopedia,* Available at http://en.wikipedia.org/wiki/F1_score, last accessed 13th September 2009.

25. F. Colace, P. Foggia and G. Percannella, "A probabilistic framework for TV-news story detection and classification," *IEEE International Conference of Multimedia and Expo (iCME'05)*, pp 1350—1355, July 2005.

26. G. Harit, S. Chaudhury and H. Ghosh, "Using Multimedia Ontology for generating conceptual annotations and hyperlinks in video collections," *International conference on Web Intelligence*, Hong Kong, December 2006

27. Anonymous, "News Ticker," *Wikipedia – The free Encyclopedia,* http://en.wikipedia.org/wiki/News_ticker, Last accessed 13th September, 2009.

28. Dave Winer, "RSS 2.0 Specification", available at http://cyber.law.harvard.edu/rss/rss.html, Last accessed on 13th September 2009.

29. S. Kopparapu, A. Srivastava, P. V. S. Rao, "Minimal Parsing Key Concept Based Question Answering System", *HCI (3)*, 2007.

30. P. Gelin and C. J. Wellekens, "Keyword spotting for video soundtrack indexing", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 1, Page(s):299 – 302, 7-10 May 1996.

31. Y. Oh, Jeong-Sik-Park and Kyung-Mi Park, "Keyword spotting in Broadcast News ", *IGNOIE (Sendai, Japan)*, pp.208-213, January 2007**.**

32. G. Quenot, T. P. Tan, Le Viet Bac, S. Ayache, L. Besacier and P. Mulhem, "Content-Based Search in Multi-Lingual Audiovisual Documents using the International Phonetic Alphabet", *CBMI 2009*, June 3-5, Chania, Greece, 2009.

33. D. Dimitriadis, A. Metallinou, I. Konstantinou, G. Goumas, P. Maragos and N. Koziris. "GRIDNEWS: A Distribured Automatic Greek Broadcast Transcription System", *ICASSP* 2009.

34. M. Laxminarayana and S. Kopparapu, “Semi-Automatic Generation of Pronunciation Dictionary for Proper Names: An Optimization Approach”, *Proceedings of ICON-2008: 6th International Conference on Natural Language Processing*, CDAC, Pune, India, pp. 118-126, Dec 20-22 2008.

35. J. Yi, Y. Peng and J. Xiao, “Color-based Clustering for Text Detection and Extraction in Image”, *MM’07*, September 23–28, Augsburg, Bavaria, Germany, 2007.

36. J. Sun, Z. Wang, H. Yu, F. Nishino, Y. Katsuyama and S. Naoi, "Effective Text Extraction and Recognition for WWW Images", *DocEng ’03*, Grenoble, France, November 20-22, 2003.

37. Q. Yea, Q. Huang, W. Gao, D. Zhaoc, "Fast and robust text detection in images and video frames", Elsevier, Image and Vision Computing 23, 2005.

38. J. Gllavata, R. Ewerth and B. Freisleben, "Tracking Text in MPEG Videos", *ACM,* 2004.

39. A. D. Bagdanov, L. Ballan, M. Bertini, and A. D. Bimbo, "Trademark Matching and Retrieval in Sports Video Databases", MIR’07, Augsburg, Bavaria, Germany, September 28–29, 2007.

40. R.Y. Tsai and T.S. Huang, “Multiple frame image restoration and registration,” *in Advances in Computer Vision and Image Processing,* Greenwich, CT: JAI Press Inc., pp. 317-339, 1984.

41. V. H. Patil, “Color Super Resolution Image Reconstruction,” *International Conference on Computational Intelligence and Multimedia Applications*, 2007.

42. P. Vandewalle, S. S¨usstrunk, and M. Vetterli, “Lcav super-resolution source code and images“, available at http://lcavwww.epfl.ch/reproducibleresearch/ VandewalleSV05, Last accessed on 13th September 2009.

43. N. Otsu, "A threshold selection method from gray-level histograms". *IEEE Trans. Sys., Man., Cyber.* **9**: 62–66, 1979.

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

45. Gurpreet Singh Lehal, "Optical character recognition of Gurmukhi script using multiple classifiers", Proceedings of the International Workshop on Multilingual OCR, Barcelona, Spain,July 2009.

46. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, "A Bilingual OCR for Hindi-Telugu Documents and its Applications," *ICDAR, vol. 1, pp.408, Seventh International Conference on Document Analysis and Recognition (ICDAR'03)* - Volume 1, 2003.

47. S.V. Rice, F. R. Jenkins, and T. A. Nartker**, “**The Fourth Annual Test of OCR Accuracy**”,** *Technical Report 95-04, Information Science Research Institute, University of Nevada*, Las Vegas, April 1995.

48. R. Smith, “An Overview of the Tesseract OCR Engine” *Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR* Volume: 2, pp: 629-633, 23-26 Sept. 2007.

49. Anonymous. “Levenshtein distance”. Wikipedia, the free encyclopedia. Available at http://en.wikipedia.org/wiki/Levenshtein_distance. Last accessed 13th Sepetember, 2009.

50. K. Vaiapury, P. K. Atrey, M. S. Kankanhalli, K. Ramakrishnan, “Non-Identical Duplicate Video Detection Using the SIFT Method”, *In IET Conference on Visual Information Engineering*, 2006.

51. Y. Ke., R. Sukhtankar, “PCA-SIFT: A More Distinctive Representation for Local Image Descriptors”, *In Proc. Of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2004.*

52. J. Haitsma, T. Kalker, “A Highly Robust Audio Fingerprinting System”, *Proceedings of International Symposium on Music Information Retrieval,* 2002.