INDEXING OF MULTILINGUAL NEWS TELECAST USING AUDIO-VISUAL KEYWORDS

H. Ghosh, A. Khare, A. Gorai

TCS Innovation Labs Delhi
TCS Towers, 249 D&E Udyog Vihar Ph-IV
Gurgaon 122015, INDIA

S. K. Kopparapu, M. Pandharipande

TCS Innovation Labs Mumbai
Yantra Park, Pokran Road No. 2
Thane West 400601, INDIA

ABSTRACT

Indexing of news video streams with semantic keywords is of interest to agencies that regularly monitor many news channels. In this paper, we describe a new method for indexing news video in different languages, for which there are inadequate language tools. Our approach involves combining multimodal inputs, namely audio and visual, and spotting of a handful of keywords with higher reliability, as compared to creating complete transcripts. We conduct a set of experiments to establish performance improvement despite lesser reliability of the language tools.

***Index Terms*—** TV broadcasting, Optical character recognition, Speech recognition, Information retrieval

1. INTRODUCTION

Indexing of telecast video streams with semantic keywords is important for several agencies, who need to monitor a number of news channels to track events in specific domains, such as sports and politics. In this paper, we propose a novel method for indexing news video in English and different Indian languages with keywords spotted in audio and visual components of the video stream. The major challenge faced by us has been the unavailability of closed-captioned-text with Indian channels and lack of reliable language tools for Indian languages. Indexing of news broadcasts in different languages based on textual transcripts of speech has been reported in [1, 2]. There are a few commercial tools for speech transcriptions available in English and handful of other European and Asian languages. Speech recognition tools for Indian languages are not robust enough for creating reliable transcripts. Moreover, varied accents of news readers from different parts of India makes transcription of even the English channels unreliable. A complementary approach is to create transcripts of text presented in visual form [3, 4]. Use of non-standard fonts in many Indian languages make OCR unreliable.

Our challenge has been to provide reliable indexing of Indian transmissions despite inadequacies of the language tools. Considering that the users of the system are generally interested in a specific domain, we redefine the goal of our system as indexing the news telecast with a handful of keywords characterizing the domain of interest, rather than attempting a complete transcript of the speech. Spotting a few keywords in speech is more robust than creating a complete transcript. Further, we enhance the indexing performance of the system by adding the keywords spotted in the ticker text through visual processing. While use of a restricted set of keywords enhances the system performance, selection of domain keywords is a non-trivial task. We derive the keyword list of contemporary interest by analyzing the RSS (Really Simple Syndication is a family of web feed formats used to publish frequently updated works) feeds in the domain of user interest. This technique makes the keyword list succinct as well as up-to-date. While only English RSS feeds are processed for this purpose, the identified keywords are either translated or transliterated to different Indian languages for keyword spotting in those languages.

The rest of the paper is organized as follows. Section 2 provides a description of the system, including methods used to derive the keyword list and keyword spotting in audio and visual components of news video. Section 3 provides experimental results for indexing performance. Finally, section 4 concludes the paper.

2. SYSTEM DESCRIPTION

Figure 1 provides an overall system architecture. News videos are recorded from the telecast stream in small pieces. The recorded video files are then subjected to audio and visual keyword spotting techniques. The results from the two processes are combined to form a consolidated list that is used to index the video. The list of keywords to be spotted is obtained from a public RSS feed.

2.1. Creating the keyword list

RSS feeds provide the headlines and links to contemporary news stories. Users are generally interested about specific events, persons or places. We select the common and the proper nouns from the RSS feed and the associated stories to populate the keyword list, using a named entity detection

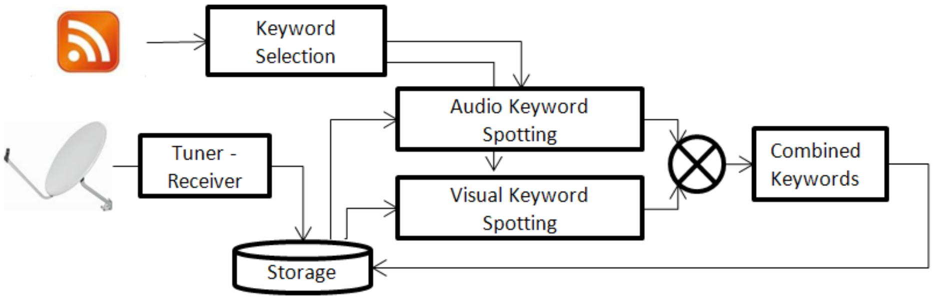

Fig. 1. System Architecture

module [5] and frequency count. RSS feeds generally categorize the news items in several categories. Keyword selection is restricted to news items in a few specific categories of user interest to keep the keyword list short.

The English keywords so derived are converted into different Indian languages. The proper nouns are transliterated using a pronunciation lexicon [6]. The nouns in Indian languages are generally phonetic and their transliteration is significantly simpler than that for English and other European languages. The common nouns need to be translated from English to different Indian languages. In general, there are *many-to-many* mapping between English and Indian language keywords. We propose to use a multilingual dictionary [7], that establishes equivalence between groups of words (*synsets*) in different Indian languages using English as the pivotal language.

The multilingual keyword list so created is used for keyword spotting in both audio and visual components of recorded newscasts. A significant advantage of obtaining a keyword list from the RSS feed is the currency of the keywords because of dynamic updates of the RSS feeds. While there are RSS feeds in some Indian languages, aligning the words in different languages from independent sources is significantly difficult compared to our approach discussed above.

Figure 2 depicts some example entries in a multilingual keyword list in English and Bangla, an important Indian language. While the first two entries in the figure represent proper nouns (names of a person and a country), the third entry corresponds to a common noun and has two Bangla synonyms.

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

        <CONCEPT NAME=”Afghanistan”>
             <ENG KEY=”Afghanistan”>Afghanistan</ENG>
             <BEN KEY=”Afganistan”>আফগানিস্তান</BEN>
        </CONCEPT>

        <CONCEPT NAME=”Rajshekhar”>
             <ENG KEY="Rajshekhar">Rajshekhar</ENG>
             <BEN KEY=”Rajshekhar”>রাজশেখর</BEN>
        </CONCEPT>

        <CONCEPT NAME=”Terrorist”>
             <ENG KEY=”Terrorist”>Terror</ENG>
             <BEN KEY=”Santraas Aatank”>সন্ত্রাস আতঙ্ক</BEN>
        </CONCEPT>
   </L>
</RULE NAME>
```

Fig. 2. Multilingual keyword list

2.2. Keyword spotting in speech

Audio keyword spotting system essentially enables identification of spoken words or phrases of interest in an audio broadcast.

Most of the audio keyword spotting systems take an acoustic speech signal, a time sequence, $x(t)$, as input and use a set of N keywords ($\{K_i\}_{i=1...N}$), as reference to spot the occurrences of these keywords in the speech signal [8]. A speech recognition engine can be looked upon as $S : x(t) \rightarrow p(s)$, where $p(s)$ is a phoneme string sequence $\{p_k\}_{k=1...M}$. Internally, the speech recognition engine has a built in pronunciation lexicon which is used to associate the entries in the keyword list with an equivalent phonemic sequence. The recognized phonemic string from the acoustic audio, $\{p_k\}_{k=1...M}$, is matched with the keywords (phonemic equivalent of keywords).

If any part p_k sequence matches any of the N keywords, namely $\{K_i\}_{i=1...N}$, then S, is deemed to have spotted a keyword.

A functional keyword spotting system is shown in Figure 3. The first step is to extract the audio track from a video news recording. The keyword list described Section 2.1 is converted into a speech grammar file used by the speech recognition engine [6]. Acoustic models and the speech grammar file are then used to mark all possible occurrences of the keywords in the acoustic stream. The output is typically the spotted keywords and the time instance at which that particular keyword occurred.

We propose a multi-lingual audio KWS system that can be

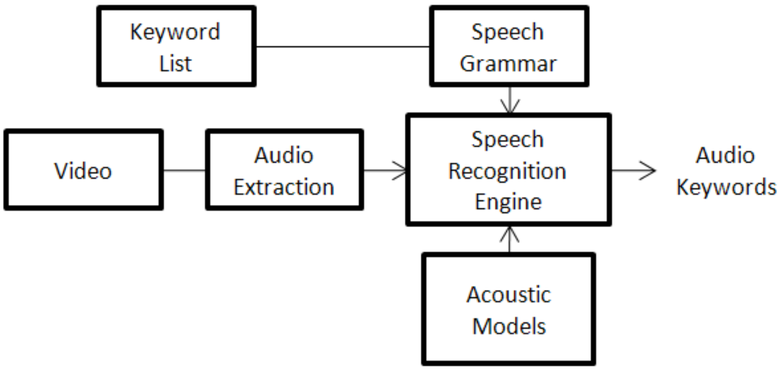

Fig. 3. Keyword spotting in speech

used to spot keywords in English and Indian Language newscasts. The grammar file, or the pronunciation lexicons, for the common nouns can be readily derived from conventional pronunciation dictionaries of the different languages. However, creation of pronunciation lexicon for proper nouns [6] is not so trivial. We take advantage of the fact that Indian proper names, are pronounced in the same way in different Indian languages and in that sense are language independent. This implies that the same grammar file for keywords representing proper nouns can be used irrespective of language of broadcast.

2.3. Keyword spotting in ticker text

Ticker text refers to a small, typically 10-15%, of screen space dedicated to presenting information in textual form. The static ticker text band generally contain more information value than the scrolling ones. Figure 4 shows the steps involved in extracting text from a static ticker text band. The first stage of processing involves extraction of ticker text regions from the video frames. A ticker-text band generally occupy a fixed position in a news channels. We create a meta-information file that contains the ticker text region definition in the video frames for every channel. Moreover, it is necessary to know the language of transmission for the channel to subject the ticker text images to a suitable OCR tool. This information is also encoded in the meta-information file. The extracted ticker text regions are binarized using Otsu's method [9].

Ticker texts are generally stable for a small time duration, so that it is possible for a human being to read it. The next stage of processing involves discovering a sequential group of ticker text images containing the same text. This is done by comparing successive ticker text images much in the same way as shot detection in videos. We call the resultant groups stable ticker text groups.

Let $p_{i,x,y}$ represent the pixel in y^{th} row and x^{th} column of the ticker text image T_i for i^{th} frame. For binary images, $p_{i,x,y} \in 0, 1$. We define the distance between i^{th} and $i + 1^{th}$ frames as

$$d_i = \sum_{x,y} |p_{i+1,x,y} - p_{i,x,y}|$$

The two frames are assumed to contain same ticker text if $d_i < \delta$, where δ is a threshold. If $d_{i-1}, d_{i+n} \geq \delta$ and $\forall k_{(0 \leq k < n)}, d_{i+k} < \delta$, the set $\mathcal{T} = \{T_i, T_{i+1}, \cdots T_{i+n}\}$ forms a maximal ticker text group. $\mathcal{T}$ is considered to be a stable ticker text group, if n is greater than a certain number of frames. Note that a stable ticker text groups contain several images, with identical text.

We choose four arbitrary images from stable ticker text groups $\mathcal{T}$ to create an enhanced image using image superresolution method [10] which improves the performance of OCR processing.

We have used Tesseract OCR engine [11], which supports English and several Indian languages, e.g. Bangla [12], to extract the text from the super resolution ticker text images. The final stage of processing involves keyword spotting on the OCR output. While the OCR output is fairly accurate, we accept close matching words using weighted Levenshtein distance. The weights are selected based on visual similarity of letters, e.g. *l* (twelfth character in English alphabet in small) and *1* (number one) have a small weight separating them.

3. RESULTS

We have tested the performance of keyword based indexing with several news video snippets recorded from different Indian channels in English and in Bangla (one of the major Indian languages), over two consecutive days. Our data-set comprised 9 English and 5 Bangla news clips, each of which have a duration between 40 and 120 sec. We have used RSS feeds from "Headlines India" over the same days for creating the keyword list. We studied (a) the performance improvement of the system with the use of a restricted keyword list and (b) improvement due to combined audio-visual cues. In the following experiments, the ground truth, i.e. the actual occurrence of a keyword in the speech or visual component of a

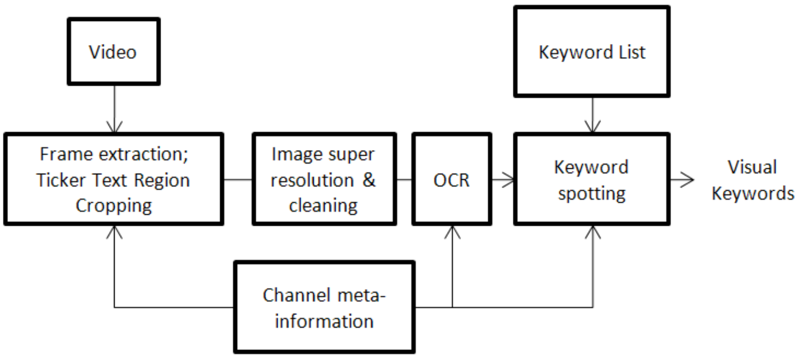
Video
Keyword List
Frame extraction; Ticker Text Region Cropping
Image super resolution & cleaning
OCR
Keyword spotting
Visual Keywords
Channel meta-information

Fig. 4. Keyword spotting in ticker text

video is established with manual inspection of the videos.

For studying the effect of the length of the keyword list on the system performance, we used the standard retrieval performance metrics *recall* and *precision*. Let, n denote the number of instances of any of the keywords being present in a news clip, k denote the number of keywords reported and r denote the number of such keywords correctly identified. Then recall ($\mathcal{R}$) and precision ($\mathcal{P}$) parameters are defined as $\mathcal{R} = \frac{r}{n}$, and $\mathcal{P} = \frac{r}{k}$.

We created (a) a master keyword list by considering all stories from the RSS feed and (b) a restricted keyword list by considering only “India news” category. The lists contain 137 and 16 English keyword groups (mostly proper nouns) respectively and their Bangla equivalents. Table 1 depicts the average recall and precision values for English and Bangla channels with master and restricted keyword lists for keyword spotting in speech and ticket text. We note that restricting the keyword list improves the results (both recall and precision), more significantly for speech processing. We find the improvement to be significantly larger for Bangla (both for speech and OCR), where the language tools are less mature, than in English. The performance of OCR has been found to be significantly higher than keyword spotting in audio. This is likely to be caused by exploitation of the redundancy in the groups of ticker text images carrying the same text.

Next, we study the impact of combining audio and visual cues for indexing. This experiment has been done with restricted keyword list only. The goal of the system is to index the videos with as many keywords from the keyword list as possible. Thus, we define the indexing performance of the system as:

$$IP = \frac{|k|}{|K|} \times 100\%$$

where, K is the set of distinct keywords actually present in the video and k is the set of distinct keywords correctly identified (and used for indexing). Obviously, $k \subseteq K$.

We measure the indexing performance of the system using audio and visual keyword spotting individually and also by combining the two methods. Let K_a and K_v denote the set of distinct keywords actually occurring in the speech and the visuals respectively in a section of the news video. Then, $K_o = K_a \cup K_v$ represents the set of all keywords appearing in the video segment. 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 respectively) can be measured as

$$IP_a = \frac{|k_a|}{|K_a|} \times 100\%, IP_v = \frac{|k_v|}{|K_v|} \times 100\%$$

$$IP_o = \frac{|k_a \cup k_v|}{|K_a \cup K_v|} \times 100\%$$

	Master list (137 word groups)		Restricted list (16 word groups)	
	Recall	Precision	Recall	Precision
Keyword spotting in audio				
English	14.63	47.00	25.55	72.32
Bangla	10.53	28.57	25.00	73.08
Overall	13.99	42.97	25.45	72.46
Keyword spotting in ticker text				
English	74.19	96.98	77.03	100.00
Bangla	62.50	93.54	70.59	100.00
Overall	72.57	95.75	76.17	100.00

Table 1. Recall and Precision values with Master and Restricted keyword lists

Table 2 depicts the indexing performance of the audio, the visual and the overall system. Note that the indexing performance is significantly higher than the corresponding recall values because of redundancies in the news. If any one occurrence of a keyword, either in speech or in visual, is spotted, it is considered as a success. The overall indexing performance for the videos in both the languages, English and Bangla, is significantly higher than when any one of the modes, audio or

	Audio (IP_a)	Visual (IP_v)	Combined (IP_o)
English	64.71	79.12	86.55
Bangla	63.33	72.73	82.50
Overall	64.45	77.88	85.53

Table 2. Indexing performance with multimodal cues

visual, has been used. This is expected since redundancy in multimodal cues improves indexing performance.

4. CONCLUSION

Our primary contribution in this paper has been reliable indexing of Indian news telecasts with keywords despite inadequacies in the available language tools. The main factors behind this reliable indexing has been the selection of a restricted domain-specific keyword list, in contrast to attempting a complete transcription. Use of several image and audio processing stages has also added to the reliability of system. Use of RSS feeds to identify the keywords of interest results in automatic regular update of the keyword list and contemporariness of the system. The conversion of English keywords to their Indian Language equivalents helps indexing Indian language transmissions with English keywords.

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. A particular challenge is the languages, for which language tools do not exist and are unlikely to be available in foreseeable future. We propose to use image and audio processing methods for comparing them with already indexed videos.

5. REFERENCES

[1] J. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," in *Communications of the ACM*, February 2000, vol. 43, pp. 64–70.

[2] Helen M. Meng, Xiaoou Tang, Pui Yu Hui, Xinbo Gao, and Yuk Chi Li, "Speech retrieval with video parsing for television news," in *Programs, Proceedings of ICASSP*, 2001, pp. 1401–1404.

[3] D. Crandall, S. Antani, and R. Kasturi, "Extraction of special effects caption text events from digital video," in *International Journal on Document Analysis and Recognition*, 2003, vol. 5, pp. 138–157.

[4] H. Li, D.S. Doermann, S. David S., and O.E. Kia, "Text extraction, enhancement and ocr in digital video," in *Selected Papers from the Third IAPR Workshop on Document analysis Systems*, 1999, pp. 363–377.

[5] S. Kopparapu, A. Srivastava, and P. Rao, ," in *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, Julie Jacko, Ed. 2007, vol. 4552 of *Lecture Notes in Computer Science*, pp. 104–113, Springer Berlin / Heidelberg.

[6] M. Laxminarayana and S. Kopparapu, "Semi automatic generation of pronunciation dictionary for proper names an optimization approach," in *Proceedings of International Conference on Natural Language Processing*, 2008, pp. 118–126.

[7] R. K. Mohanty, P. Bhattacharyya, S. Kalele, P. Pandey, A. Sharma, and M. Kopra, "Synset based multilingual dictionary: insights, applications and challenges," in *GWC 2008:The Fourth Global WordNet conference*, January 2008, pp. 321–332.

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

[9] N. Otsu, "A threshold selection method from gray-level histogram," in *IEEE Trans on Systems, Man and Cybernetics*, January 1979, vol. SMC-9, pp. 62–66.

[10] P. Vandewalle, S. Susstrunk, and M.Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," in *EURASIP Journal on Applied Signal Processing*, 2006.

[11] R. Smith, "An overview of the tesseract ocr engine," in *Proc. 9th International Conference on Document Analysis and Recognition*, 2007, vol. 2, pp. 629–633.

[12] 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.