# Creation and Analysis of a Corpus of Text Rich Indian TV Videos

T. Chattopadhyay, Soumik Sengupta, Aniruddha Sinha
*Innovation Lab*
*Tata Consultancy Services*
*Kolkata, India*
*Email: {t.chattopadhyay, soumik.sengupta, aniruddha.s}@tcs.com*

Nisha Rampuria
*Trainee in Innovation Lab*
*Tata Consultancy Services*
*Kolkata, India*
*Email: n.rampuria@yahoo.com*

***Abstract*—A lot of research is now going on to extract the context of the show to provide additional information related to the TV show. One major method to extract the context from TV is to recognize the texts from the videos which is also known as video Optical Character Recognition (VOCR). The problem of VOCR from the TV shows of a multiligual country like India is more difficult. In India still more than 90% TV viewers are using RF Cable as input to TV and nearly 90% channels have multilingual texts in the TV shows. Thus the video quality is poor in compare to the modern digital TV signals as well as different text scripts are present in a single video frame. These made the problem of Indian TV context recognition more challenging. So this paper is concerned about the construction of a video corpus of text rich Indian TV shows. The proposed database contains more than 100 videos each of nearly 10 min duration containing text in the video frame. A statistical analysis of the corpus is also presented in the paper which can be used to identify the genre of TV show. The analysis also revealed that distribution of numerals, special characters, uppercase and lower case character can be used to classify a news video frame. This corpus is useful for a wide variety of research problems namely, (i) localization of the text regions from a video frame, (ii) recognition of texts from a video frame, (iii) extraction of context from video, and (iv) performance evaluation of a video OCR system.**

***Keywords*-Video OCR; Corpus; Indian TV video; Indian TV Video Analysis;**

## I. Introduction

Recent market trends on consumer electronics shows that the demand for Internet connected TV is large as the sale for such product raises to in the second quarter of 2009 compared to the first quarter of the same year [1]. The survey on the wish list of the customers of connected TV shows that there is a demand of a service where the user can get some additional information, from Internet. This market demand motivates the research on extracting the context from TV videos which in turn, largely depends on the extraction of the textual information from TV video. Understanding the basic TV context is quite simple for digital TV broadcast (cable or satellite) using meta-data provided in the digital TV stream. But, in developing countries, digital TV penetration is quite low. For example, in India, more than 90% TV households still have analog broadcast cable TV. Understanding the TV context in the analog broadcast scenario is really a big challenge. So lots of research is now going on to recognize the text from video. This research needs to (i) localize the text region from video frame and then (ii) recognize the text within it to get the textual context of the video frame. The work presented in this paper is motivated to facilitate such research on text recognition from TV videos. Moreover, unavailability of suitable corpora of such TV videos has so far prompted the researchers to define their own data set for testing their algorithms. Even the video corpus usually referred in the literature has the following limitations:

- The video corpus available in the literature are mainly motivated to provide a common teststream for research on (i) video shot boundary detection [2] or (ii) annotate a news video automatically. But the research on video OCR evaluate both localization and recognition algorithm. As a result, replication of experiments and comparison of performance among different methods have become difficult tasks.
- While we are developing one such video OCR like solution [3], [4] we found that the video quality of the Indian TV shows are mostly having a maximum resolution of 720x576 pixels which is much less than the HD resolution which is common in developed countries like US. So the performance of video OCR is not that much impressive in Indian scenario even if the performance is pretty good for HD TV videos.
- India is a multi lingual country. So most of the regional channels consists of at least two language scripts namely English and the regional language. This makes the task of text localization and recognition more complex.
- No video corpus, to the best of our knowledge provides a annotated data to evaluate the performance of text localization module. But this module plays an important role in the research of video OCR as it has a huge contribution on the overall performance of the system in terms of accuracy and speed of execution.
  So the proposed corpus that is available on request will substantially contribute to this end. Moreover a significant statistical analysis on the annotated corpus is also presented in this paper.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.174

## II. Data Acquisition

The videos are recorded on the DSP platform in YUV format. There are many ways to record a TV shows are there. Some of them and their limitations are like:

- Personal Video Recorder (PVR) of a Set Top Box (STB) may be an option to record videos from TV. The main problem of videos recorded by PVR is that they are in compressed format. So some loss in video quality is incurred during encoding. Moreover the compressed videos first needs to be decoded and then the algorithms are needed to be applied. Thus the overall complexity of the system is high. The compressed domain methods are not that much accurate in compare to the pixel domain approach.
- USB TV tuner card may also be another alternative. But they also record the video in compressed format and thus the above mentioned problems still exists for this method, too.
- YouTube also includes different videos on TV shows. But the main problem is also the inferior video quality and loss in video quality because of multiple data compression.

So we have used a modified version of the Home Infotainment platform (HIP) [5] to record the videos from TV in raw YUYV format. We have observed that the texts in a video usually persists in the screen for more than 2-3 seconds. So it is not required to process every frame in the research of video OCR. So we have recorded 5 frames per second so that even the issue of illegal distribution of those TV contents can also be avoided.

The method for recording is presented very precisely here:

Broadcast TV content is fed into the video capture module via analog tuner or composite "Video In" port which then enters the post processing sub-system via the context analysis interface. The video input coming to the hardware is captured using Linux V4L2 interface as analog video input sources. The capture application makes "V4L2 IOCTL" calls to configure the capture driver and de-queue frames from capture array. Once the frames are copied to user space for post processing, the original frame is en-queued back to be overwritten. This is explained in Figure 1.

## III. Selection of TV shows

We have observed that some genres of TV videos contains rich textual contents. We have identified five such categories of videos where textual content is very high like:

- *News videos:* News videos are containing lots of textual information like breaking news, ticker text, stock ticker. Some possible applications might be duplicate news story detection, personalized stock market ticker, Personalized mash-up of the internet news with TV news, etc. One such example of news video is shown in Fig. 2.

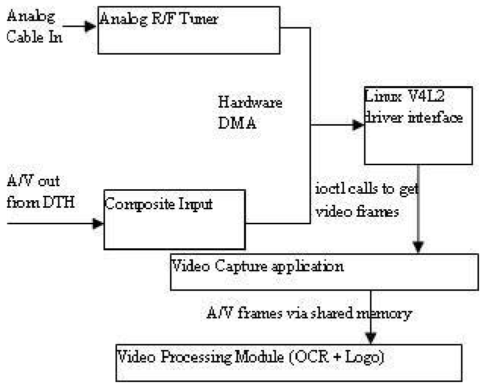
Analog Cable In
Analog R/F Tuner
Hardware DMA
Linux V4L2 driver interface
A/V out from DTH
Composite Input
ioctl calls to get video frames
Video Capture application
A/V frames via shared memory
Video Processing Module (OCR + Logo)

Figure 1. Data Acquisition System Overview

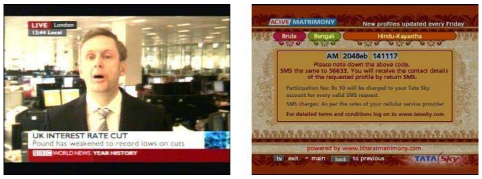

Figure 2. News Video and Active Page

- *Active pages and Guide pages of DTH services :* Active Pages usually contain an alpha-numeric stream that needs to be sent to a particular subscriber number to avail the service like downloading wallpaper, ring tone, or any other service. As the existing DTH services lack interactivity the user need to manually type the content and send to the specific recipient number as short messaging service (SMS). We have recognized the text from those active pages automatically so that any application can be built to send those text automatically as an SMS. One such example is shown in Fig. 2.
- *Recipe Shows:* One of the popular shows in India particularly for women are recipe show. Recipe shows usually contains the ingredients and recipe as a text at the end of the show. Many applications can be developed like sending automatic SMS containing recipe to intended user. One such examples are shown in Fig. 3.
- *Sports Videos:* Sports videos contains different textual information like score, player names. One such example is shown in Fig. 4.
- *Movies and Music with Subtitles:* Music channels usually contain the album name, the singer name as a text during the music video. These textual information can be used to build different applications. Movies also contains subtitle. One such example is shown in Fig. 4.

The proposed database contains nearly 1000 Minutes of video containing textual information. We have observed that out of the different genres of shows telecasted in TV, some

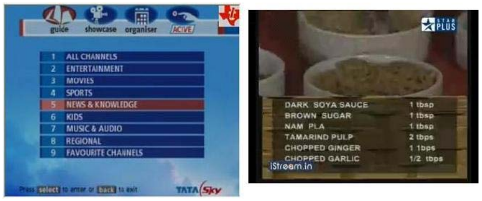

Figure 3. Guide Page and Recipe Video

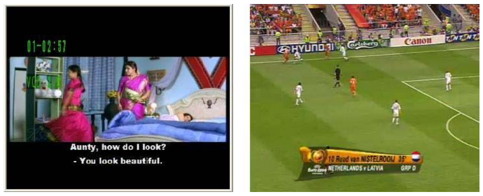

Figure 4. Video with Subtitle and Sports Video

types of shows are text intensive and textual contexts can be extracted from them to build an application. This list is presented in Table I.

## IV. Statistical Analysis of the Corpus

The proposed database contains more than 100 Indian TV shows and 996 minute duration. Genre wise distribution of the recorded video is presented in Table II. The text rich TV shows were recorded from (i) News shows which contains lots of text information like stock ticker, ticker news, breaking news, channel information, date and time of the show. (ii) Music videos which also contains a lot of meaningful data in text format like music video name, album name, singer name, next show information. (iii) Sports videos which gives the information about the score at that instant of time, competitive teams names, sponsor information, etc. TV shows were recorded from TV shows telecasted in West Bengal, a state in eastern India. We have made different analysis on the recorded videos.

### A. Genre Wise Analysis

The average number of character and words present in a video frame is varying largely for different genres as shown in Table III. So any heuristic can be applied to classify the genres of videos based on the average number of character present in a video frame. This average value ($\mu$) obtained for each genre can be used as a threshold value for this feature. The tolerance factor can be computed from the standard deviation ($\sigma$) of character per frame for different genre. It can be found that the scope of false classification for sports and music videos is also very less if we use $\sigma$ as the tolerance factor along with $\mu$. The values of standard deviation for different genres are shown in Table IV.

Table I
List of Text Rich TV Shows

| Genre | Context | Application |
|---|---|---|
| News | Breaking News | Related web information |
| News | News text | Cross lingual information |
| Business | Stock Ticker | Personalized stock ticker |
| Music | Album and Singer name | Automatic song download |
| Movies | Subtitle | Cross lingual subtitle |
| Sports | Player name and score | Statistics of player |

Table II
Genre wise distribution

| Genre | Language | Number of channels | Duration (in Minute) |
|---|---|---|---|
| News | English | 8 | 503 |
| News | Hindi | 4 | 218 |
| News | Bengali | 1 | 57 |
| Music and movies | English | 2 | 80 |
| Music and movies | Hindi | 1 | 78 |
| Sports | English | 3 | 50 |

Table III
Genre wise average character and word length

| Genre | Char/frame | Word/frame | char/word |
|---|---|---|---|
| Sports | 27.74143401 | 4.017766497 | 6.904690461 |
| Music | 50.14246525 | 13.77414045 | 3.640333524 |
| News | 78.37634292 | 10.10357632 | 7.757287166 |

Table III shows that the text content in News shows are quite high (average 78.4) in compare to music (50.1) and sports (27.7) videos. So if we can compute the average character per frame, we can use this information to classify the TV show into different genres.

### B. Features of news video

We have observed that any news video usually contains different types of texts like news ticker, the text of the news the anchor is speaking about, breaking news of the time, date and time of the show, channel information, stock update. One example showing these different types of texts is shown in Figure 5.

The motion of the text can be used to segregate the ticker news against the static texts very easily. We have done a statistical analysis on the different types of texts and observed that the the distribution of capital case and small case alphabets, numerals and special characters can be used to distinguish between date and time, breaking news, news text and stock update.

Table IV
Genre wise standard deviation of average character and word length

| Genre<br>Genre | Standard deviation<br>Char/frame | Standard deviation<br>Word/frame |
|---|---|---|
| Sports | 10 | 1.49 |
| Music | 6.7 | 1.91 |
| News | 23.43 | 4.43 |

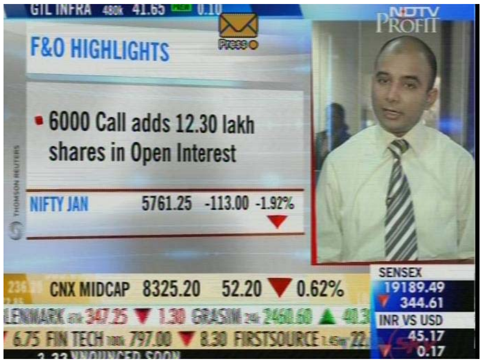
GTL INFRA 480k 41.65 0.10
NDTV PROFIT
F&O HIGHLIGHTS
6000 Call adds 12.30 lakh shares in Open Interest
NIFTY JAN 5761.25 -113.00 -1.92%
CNX MIDCAP 8325.20 52.20 0.62%
SENSEX 19189.49 344.61
INR VS USD 45.17 0.17
LENMARK 347.25 1.30 GRASIM 2460.60 40.3
6.75 FIN TECH 797.00 8.30 FIRSTSOURCE 22.

Figure 5. Stock ticker, current news in a news video

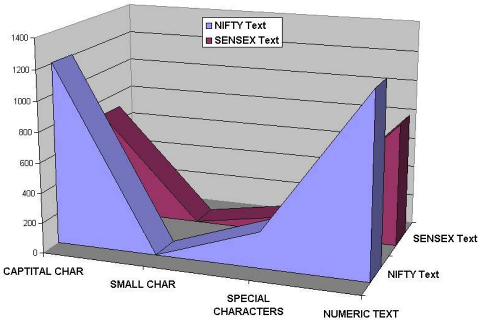

Figure 6. Distribution of characters in Stock update

*1) Stock update:* In Figure 6 we have shown the distribution of different types of character of stock update in a news show.

This figure clearly reveals that the upper case letters and numerals are most prominent in case of stock updates. It is also supported by the fact that the stock update is usually represented as a company name (in three upper case letter code) followed by the numerals showing the value of that stock.

*2) News update:* The distribution of different types of character in current news is shown in Figure 7. The figure shows that the most of the characters in this type of news update are in lower case. It is also true that in running text we usually write the first letter of the sentence or proper noun in capital letter and rest are in lower case.

*3) Date and Time:* The distribution of different types of character in date and time part displayed in a news video is shown in Figure 8. The figure shows that the most of the characters in this type of date and time are numerals and some are in lower case.

*4) Breaking News:* The distribution of different types of character in breaking news is shown in Figure 9. The figure shows that the most of the characters in this type of news update are in upper case. As the letters in upper case are more attractive to human eyes and it is a custom to make the highlighted texts in upper case, the breaking news are always coming in capital letter.

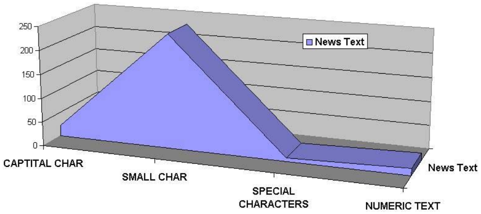

Figure 7. Distribution of characters in current news

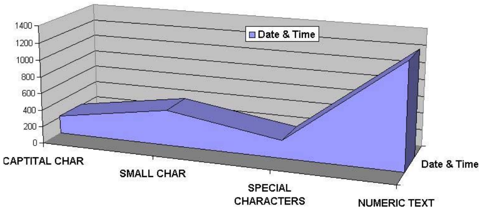

Figure 8. Distribution of characters in date and time

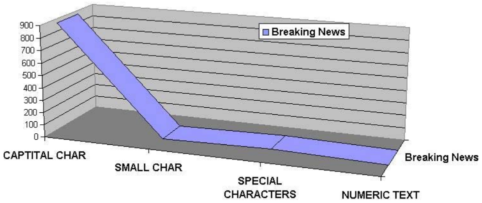

Figure 9. Distribution of characters in breaking news

So from the above discussion it is evident that the distribution of lower case (LC), upper case (UC), numerals (num), special characters (SC) within a text region can indicate the type of text in a news video. Different frequency distribution of such type of characters in different type of videos where each type is represented in percentage score, is shown in Figure 10. Also the the numerals to upper case letters shows that this ratio can also indicate the type of text in a news video. This ratio is shown in Table V.

## C. Features of music video and Sports video

In music and sports videos the distribution of different types of characters are shown in Table VI.

But some of the texts in music has very complex background and thus is difficult to segregate background and foreground from these videos. One such example is shown in Figure 11. The text regions that are difficult to recognize are marked in white circles on the screen shot.

Table V
NUMERALS TO UPPER CASE CHARACTER RATIO FOR DIFFERENT TYPES OF TEXTS IN NEWS VIDEO

| TYPE OF TEXT | % of UC | % of LC | % of SC | % of NUM | num/cap |
|---|---|---|---|---|---|
| Breaking News | 98.07909605 | 0 | 1.920903955 | 0 | 0 |
| Date & Time | 10 | 20 | 10 | 60 | 6 |
| Stock | 45.45454545 | 0 | 9.090909091 | 45.45454545 | 1 |
| News Text | 8.020152505 | 84.1503268 | 2.682461874 | 5.147058824 | 0.641765705 |

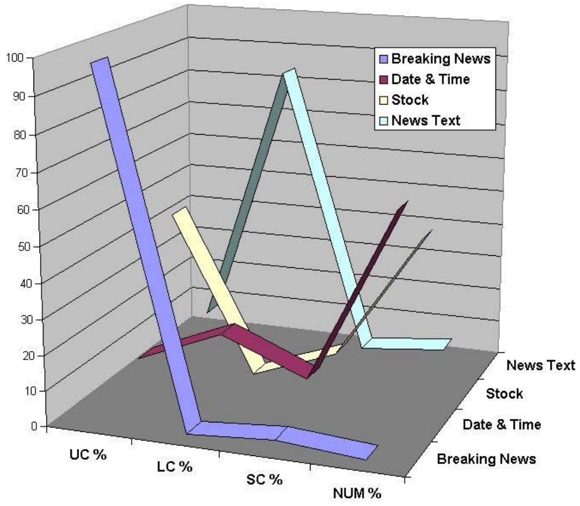

Figure 10. Percentage of occurrence of different types of character in different types of text

Table VI
NUMERALS TO UPPER CASE CHARACTER RATIO FOR DIFFERENT TYPES OF TEXTS IN SPORTS AND MUSIC VIDEO

| Genre | % of UC | % of LC | % of Num | % of SC |
|---|---|---|---|---|
| Sports | 47.11 | 16.49 | 27.42 | 10.29 |
| Music | 49.1 | 30.86 | 12.75 | 7.36 |

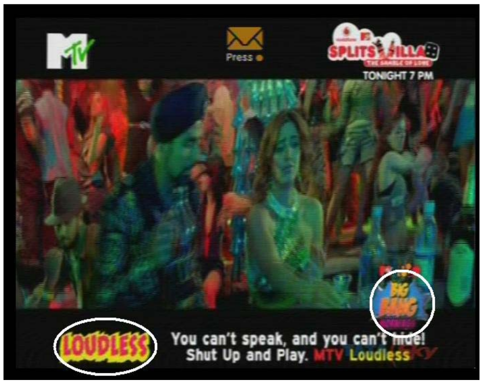
Press
SPLITS VILLA
TONIGHT 7 PM
LOUDLESS
You can't speak, and you can't hide!
Shut Up and Play. MTV Loudless

Figure 11. Complex back ground in music videos

## V. CONCLUSION

We can conclude that the corpus described in this paper addresses the following limitations of the existing corpus to facilitate the research on video OCR.

- This corpus is representative of all types of texts usually appeared on a TV show like (i) recipe show, (ii) news, (iii) sports, (iv) music, (v) movies, (vi) active pages
- It includes the videos of low resolution TV shows and low PSNR video which is common in India.
- It includes some videos in some regional languages also. So it will definitely facilitate the research on those Indian language video OCRs.
- This corpus is annotated to mark the text regions, logo regions, objects to help the research on video screen layout segmentation
- It includes high variation of text types
- A systematic analysis of texts present in a video frame is also provided that can help to formulate some heuristic to identify the intended text with the text regions of a video frame.

## ACKNOWLEDGMENT

The author would like to thank Prof B. B. Chaudhuri and Dr Utpal Garain of Indian statistical Institute, Kolkata, for their kind advice and suggestions.

## REFERENCES

[1] A. Gonsalves, *Connected TV Sales Booming*, http://www.informationweek.com/news/personal_tech/TV _theater/showArticle.jhtml?articleID=219100136, Information Week.Aug 5, 2009.

[2] T. Satou, A. Akutsu, Y. Tonomura, *Video corpus construction and analysis*,IEEE International Conference on Multimedia Computing and Systems, pp.479-485, Jul 1999.

[3] R. Lienhart,*Localizing and Segmenting Text in Images and Videos*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No. 4, April 2002.

[4] T. Chattopadhyay, A. Pal, U. Garain, *Mash up of Breaking News and Contextual Web Information: A Novel Service for Connected Television*,Proc. Of ICCCN 2010 Workshop on MCC, August 2010, Zurich.

[5] A. Pal, C. Bhaumik, M. Prashant, A. Ghose,*Home Infotainment Platform*,Proc. of International Conf. on Ubiquitous Computing and Multimedia Applications, (UCMA2010), Miyazaki, Japan, June 2010.