

Indexing of Multilingual News Videos

Sunil Kumar Kopparapu

`SunilKumar.Kopparapu@TCS.COM`

TCS Innovation Labs - Mumbai

Yantra Park, Tata Consultancy Services, Thane (West), Maharashtra, INDIA.

December 2011

What is this?

- How does one search for information?

What is this?

- How does one search for information?
- What if the information to be searched is not text?
 - Audio, Images, Videos

What is this?

- How does one search for information?
- What if the information to be searched is not text?
 - Audio, Images, Videos
- How does one enable search for information in several Indian Languages?

What is this?

- How does one search for information?
- What if the information to be searched is not text?
 - Audio, Images, Videos
- How does one enable search for information in several Indian Languages?

Digress

What is this?

- How does one search for information?
- What if the information to be searched is not text?
 - Audio, Images, Videos
- How does one enable search for information in several Indian Languages?

Digress

before that ...

What is this?

- How does one search for information?
- What if the information to be searched is not text?
 - Audio, Images, Videos
- How does one enable search for information in several Indian Languages?

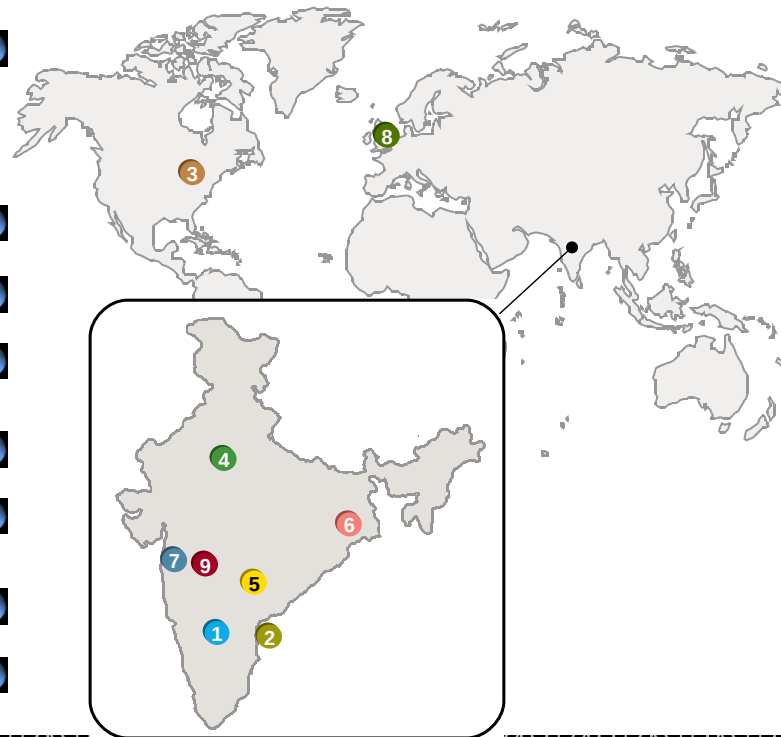
Digress

before that ...

*Wonder why is the newspapers in print media
always full?*

TCS Innovations Lab

- 1 Bangalore, India**
TCS Innovation Labs - Bangalore
- 2 Chennai, India**
TCS Innovation Labs - Chennai
TCS Innovation Labs - Retail
TCS Innovation Labs - Travel & Hospitality
TCS Innovation Labs - Insurance
TCS Innovation Labs - Web 2.0
TCS Innovation Labs - Telecom
- 3 Cincinnati, USA**
TCS Innovation Labs - Cincinnati
- 4 TCS Innovation Labs - Delhi**
- 5 Hyderabad, India**
TCS Innovation Labs - Hyderabad
TCS Innovation Labs - CMC
- 6 Kolkata, India**
TCS Innovation Labs - Kolkata
- 7 Mumbai, India**
TCS Innovation Labs - Mumbai
TCS Innovation Labs - Performance Engineering
- 8 Peterborough, UK**
TCS Innovation Labs - Peterborough
- 9 Pune, India**
TCS Innovation Labs - TRDDC - Process Engineering
TCS Innovation Labs - TRDDC - Software Engineering
TCS Innovation Labs - TRDDC - Systems Research
TCS Innovation Labs - Engineering & Industrial Services



March 10, 2010

TCS Innovations Labs - Mumbai

- Research and Development
 - Speech,
 - Script, Image
 - Natural Language Processing (Information Retrieval)
 - Wireless sensor networks
- Innovative Mobile Applications
 - mKRISHI (agro advisory system for farmers)
 - Chit Chat (audio based social networking)

TCS Innovations Labs - Mumbai

- Research and Development
 - Speech,
 - Script, Image
 - Natural Language Processing (Information Retrieval)
 - Wireless sensor networks
- Innovative Mobile Applications
 - mKRISHI (agro advisory system for farmers)
 - Chit Chat (audio based social networking)

before we go further ... a short Video

TV Broadcasting Scene

- Television Broadcasting in India
 - Main source of news and entertainment
 - Officially 22 languages
 - Television broadcast in 11 major languages (Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu)
 - Several 24×7 news channels
 - Most other non-news channels have specific news slots in their broadcast everyday

TV Broadcasting Scene

- Television Broadcasting in India
 - Main source of news and entertainment
 - Officially 22 languages
 - Television broadcast in 11 major languages (Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and Urdu)
 - Several 24×7 news channels
 - Most other non-news channels have specific news slots in their broadcast everyday

*Number of television channels growing ...
broadcast in more languages ...*

Indian News Channels

- Marathi News Channels
 - Star Majha
 - IBN Lokmat
 - Zee 24 Taas
 - Samay Mumbai

Indian News Channels

- Telugu News Channels
 - TV9, TV1, ETV2, N Tv, TV5, T News,
 - HY TV News, Gemini News, AP TV News,
 - I NEWS, HMTV, ABN Andhrajyothi, Studio N,
 - Zee 24 Ghantalu, MAHAA NEWS, Sakshi TV,
 - Channel4 News, Jagruti TV, N9 News,
 - ABC News, janata News

Indian News Channels

■ Hindi Sample!



Not bad! Just 10!

Indian News Channels

- Oops

- Aaj Tak, Sarv media group, IBN-7, Azad NEWS, News 7 Network, Khoj India, India TV, Raftaar News Channel, Live India, NDTV India, Lemon News, Channel One, India News, News 24, Press Tv, SUDARSHAN NEWS, Sahara Samay, STAR News, Zee News, Zee Business, DD News, TOTAL TV, A2Z News, Crime Nazar News, CHANNEL NO. 1, s7 News, Mahua news, ETV Bihar, Time Today, daynightnews, jansandesh.tv, GNN NEWS, P7, TV 24 News, news xpress, tv9 mumabi, Sea News, Taaza Tv, Cneb news, Khabar Bharti, VOICI - Voice Of Central India

The Scene: After Broadcast

- Post Broadcast (Indian Language)
 - Archived ... hours of news
 - Stored safely (in tapes, on hard disks!)
- Usability of the broadcast? at a later time ...
 - Can we reuse it?
 - Can we get to specific news?
 - Can we search?
- Answer
 - No

The Scene: After Broadcast

- Post Broadcast (Indian Language)
 - Archived ... hours of news
 - Stored safely (in tapes, on hard disks!)
- Usability of the broadcast? at a later time ...
 - Can we reuse it?
 - Can we get to specific news?
 - Can we search?
- Answer
 - No

But we can .. most Western news channels

The Actual Scene

- Why not for Indian channels?
 - Simple. **No** closed captioned text.
(Closed captions show a transcription of the audio portion of a program)
Unlike most Western channels!
 - No searchable text associated with news broadcast
- Who would need it anyway?
 - Security Agencies (real time monitor)
 - Individuals (Search info on media sharing sites)
 - Broadcast channels (reuse the feed as per need)

The Actual Scene

- Why not for Indian channels?
 - Simple. **No** closed captioned text.
(Closed captions show a transcription of the audio portion of a program)
Unlike most Western channels!
 - No searchable text associated with news broadcast
- Who would need it anyway?
 - Security Agencies (real time monitor)
 - Individuals (Search info on media sharing sites)
 - Broadcast channels (reuse the feed as per need)

So ... What can be done?

The Scene

- Annotate (or Index) news broadcast
- How can we do this?
 - Full transcription of the news broadcast
Good to have; difficult to automate (Poor ASR accuracies; Resource deficit)
 - Spot keywords only
Sufficient; relatively easy to automate
- For Indian channels ...
 - Need for annotating multilingual videos exists
 - Indexing can be based *only* on audio and visual cues in absence of closed caption text

The Scene

- Annotate (or Index) news broadcast
- How can we do this?
 - Full transcription of the news broadcast
Good to have; difficult to automate (Poor ASR accuracies; Resource deficit)
 - Spot keywords only
Sufficient; relatively easy to automate
- For Indian channels ...
 - Need for annotating multilingual videos exists
 - Indexing can be based *only* on audio and visual cues in absence of closed caption text

In Summary

- Searching (news) media is required and important
- Annotation of television news (or video) will help
- Difficult to annotate
 - Automatic Speech Recognition not available
 - When available - accuracies are generally very poor (ASR need to be tuned)
 - Speech resources not available for many Indian languages (speech corpus?, lexicon, dictionaries)
- Best bet is to spot selective (key) words

In Summary

- Searching (news) media is required and important
- Annotation of television news (or video) will help
- Difficult to annotate
 - Automatic Speech Recognition not available
 - When available - accuracies are generally very poor (ASR need to be tuned)
 - Speech resources not available for many Indian languages (speech corpus?, lexicon, dictionaries)
- Best bet is to spot selective (key) words

What can we do in this scenario?

The Problem

- In the absence of closed caption text ...
 - Enable indexing of multilingual videos
 - spot **only the keywords** in news broadcast
 - in different languages
 - on different television channels
 - Using
 - The audio track of the news broadcast (audio cue)
 - Ticket text ... (visual cue)
News Ticker? small screen space dedicated to presenting headlines or some important news.

The Problem

- In the absence of closed caption text ...
 - Enable indexing of multilingual videos
 - spot **only the keywords** in news broadcast
 - in different languages
 - on different television channels
 - Using
 - The audio track of the news broadcast (audio cue)
 - Ticket text ... (visual cue)
News Ticker? small screen space dedicated to presenting headlines or some important news.

Are there challenges? in keyword spotting

Challenges

- Multiple languages (22 officially recognized)
- No closed captioned text
 - Rely on audio and visual processing
- Audio or Visual to extract keywords? or Both?
 - *Maturity* of audio and visual processing work in *Indian languages* not evolved yet
 - Which keyword list to use?
 - *Construction* of keyword list in different languages?
 - Use of *small and dynamic* or *large and static* keyword list

Challenges

- Multiple languages (22 officially recognized)
- No closed captioned text
 - Rely on audio and visual processing
- Audio or Visual to extract keywords? or Both?
 - *Maturity* of audio and visual processing work in *Indian languages* not evolved yet
 - Which keyword list to use?
 - *Construction* of keyword list in different languages?
 - Use of *small and dynamic* or *large and static* keyword list

How?

Approach

- Assumption
Keyword based indexing is sufficient
- Observation
 - Keywords in news broadcast are either (a) proper nouns or (b) common nouns
 - Keywords are dynamic; they change with time
- Use RSS news feed to create dynamic keyword list (in English)
- How? Named entity detection, minimal parsing, n-gram analysis, statistical, ... several approaches

Approach

- Assumption
Keyword based indexing is sufficient
- Observation
 - Keywords in news broadcast are either (a) proper nouns or (b) common nouns
 - Keywords are dynamic; they change with time
- Use RSS news feed to create dynamic keyword list (in English)
- How? Named entity detection, minimal parsing, n-gram analysis, statistical, ... several approaches

Good for English ... for Multilingual

Multilingual Keyword list

- Identify English keyword equivalents in other Indian languages (How?)
 - Use TDIL Indian language translation tools (some language pairs exist)
 - Proper nouns - majorly independent of language (transliteration)
 - Common nouns (on-line *English to X* dictionary)
- Create a multilingual keyword list
 - input for keyword spotting in audio and
 - input for keyword spotting in visual

Multilingual Keyword list

- Identify English keyword equivalents in other Indian languages (How?)
 - Use TDIL Indian language translation tools (some language pairs exist)
 - Proper nouns - majorly independent of language (transliteration)
 - Common nouns (on-line *English to X* dictionary)
- Create a multilingual keyword list
 - input for keyword spotting in audio and
 - input for keyword spotting in visual

Multilingual keyword list? What is it?

Sample Multilingual Keyword list

```
<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">అఫఘానిస్తాన్</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">आतंकवादी</HIN>
      <TEL KEY="Atankavaadi">అతన్యవాది</TEL>
    </CONCEPT>
  </L>
</RULE NAME>
```

Sample Multilingual Keyword list

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

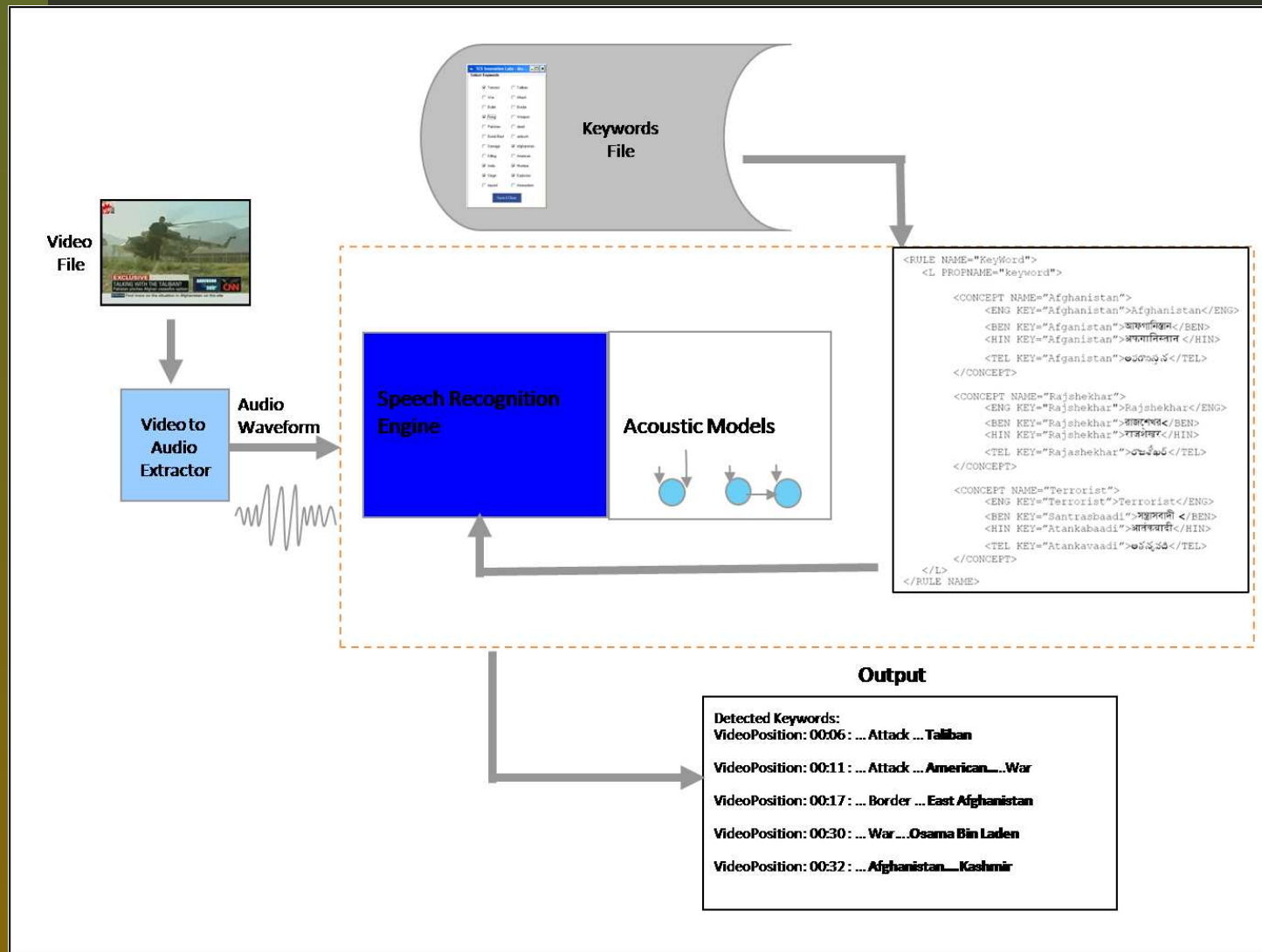
    <CONCEPT NAME="Afghanistan">
      <ENG KEY="Afghanistan">Afghanistan</ENG>
      <BEN KEY="Afganistan">আফগানিস্তান</BEN>
      <HIN KEY="Afganistan">अफगानिस्तान</HIN>
      <TEL KEY="Afganistan">అఫఘానిస్తాన్</TEL>
    </CONCEPT>

    <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">आतंकवादी</HIN>
      <TEL KEY="Atankavaadi">అతన్యవాది</TEL>
    </CONCEPT>
  </L>
</RULE NAME>
```

How is it used?

In Audio KW Spotting



Keyword spotting in audio

- Using keyword list in different languages
- Use public domain Speech recognition engine Sphinx?
- Need creation of pronunciation lexicon Indian languages phonetic
- Need language specific acoustics models different for different languages? one for all? in between?

Keyword spotting in audio

- Using keyword list in different languages
- Use public domain Speech recognition engine Sphinx?
- Need creation of pronunciation lexicon Indian languages phonetic
- Need language specific acoustics models different for different languages? one for all? in between?

Are there challenges?

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!)

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!)

Can we do something?

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!)

Can we do something?

- Build or use acoustics models for one Indian language
(simpler than building for all! Can we use existing (English) acoustic models??)
- Use this for keyword spotting
(Largely Indian language phonetic and we are doing only keyword spotting anyway!)

Challenges in Audio KWS

- Acoustic models for Indian languages do not exist
- Transcribed speech data for some languages exist (expensive and *toy like* compared to English!)

Can we do something?

- Build or use acoustics models for one Indian language
(simpler than building for all! Can we use existing (English) acoustic models??)
- Use this for keyword spotting
(Largely Indian language phonetic and we are doing only keyword spotting anyway!)

Does it work?

Some Experiments

- Used Microsoft SAPI (ASR Engine) (default English (US) acoustic models)
- Developed a self-help application based on spotting keywords
- Works well for Indian English accent (Play) (if the keywords are words not in the dictionary - we add the pronunciations)
- Works equally well with Bengali! (Play) (pronunciations lexicon for Bangla keywords words also added)

Some Experiments

- Used Microsoft SAPI (ASR Engine) (default English (US) acoustic models)
- Developed a self-help application based on spotting keywords
- Works well for Indian English accent (Play) (if the keywords are words not in the dictionary - we add the pronunciations)
- Works equally well with Bengali! (Play) (pronunciations lexicon for Bangla keywords words also added)

Definitely we can do better with acoustic models of one Indian language! Visual KW Spotting?

Keyword spotting in visual

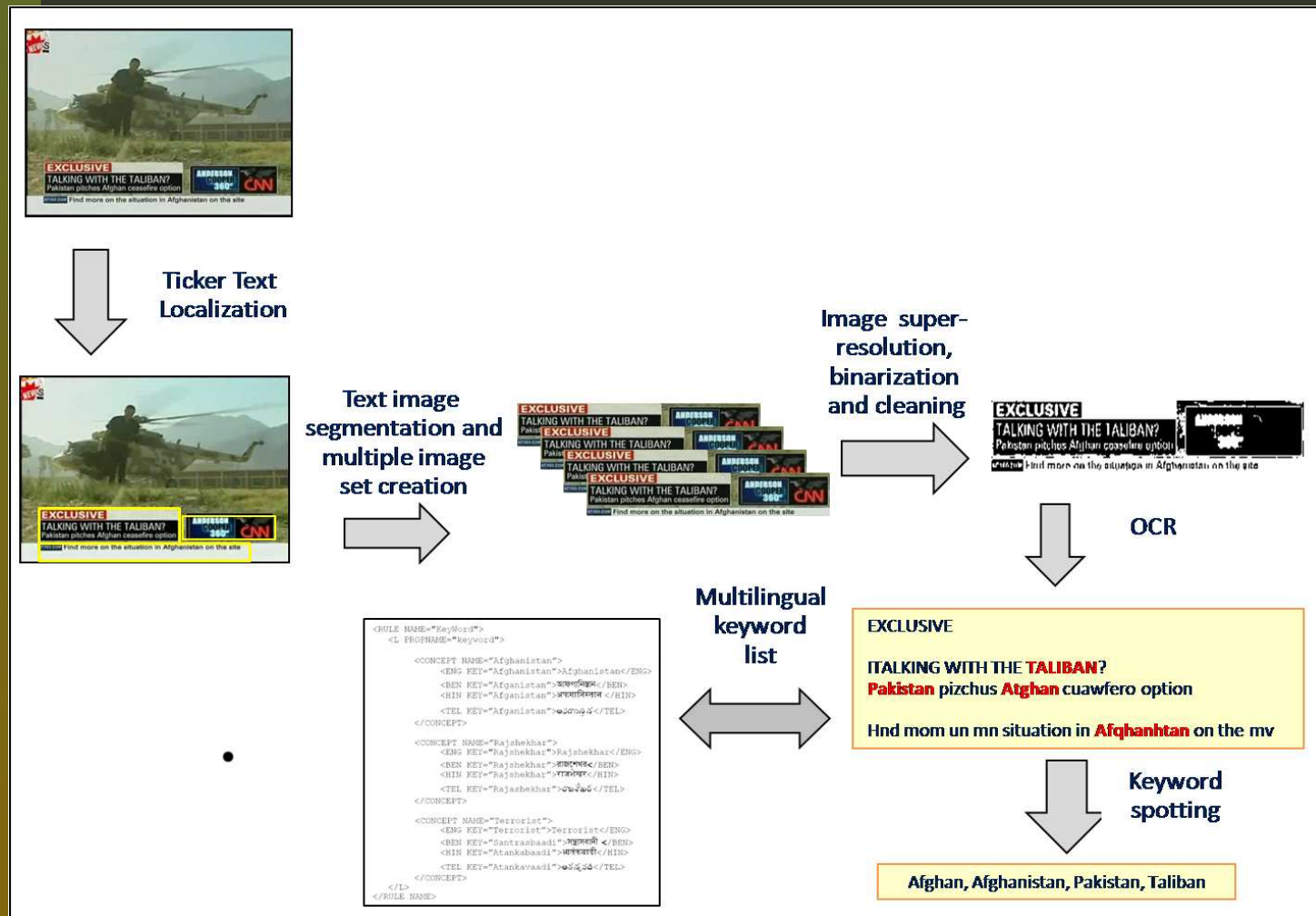
- Using keyword list in different languages
- English
 - Video OCR
 - sufficiently mature
 - increased accuracies with dynamic keyword list

Keyword spotting in visual

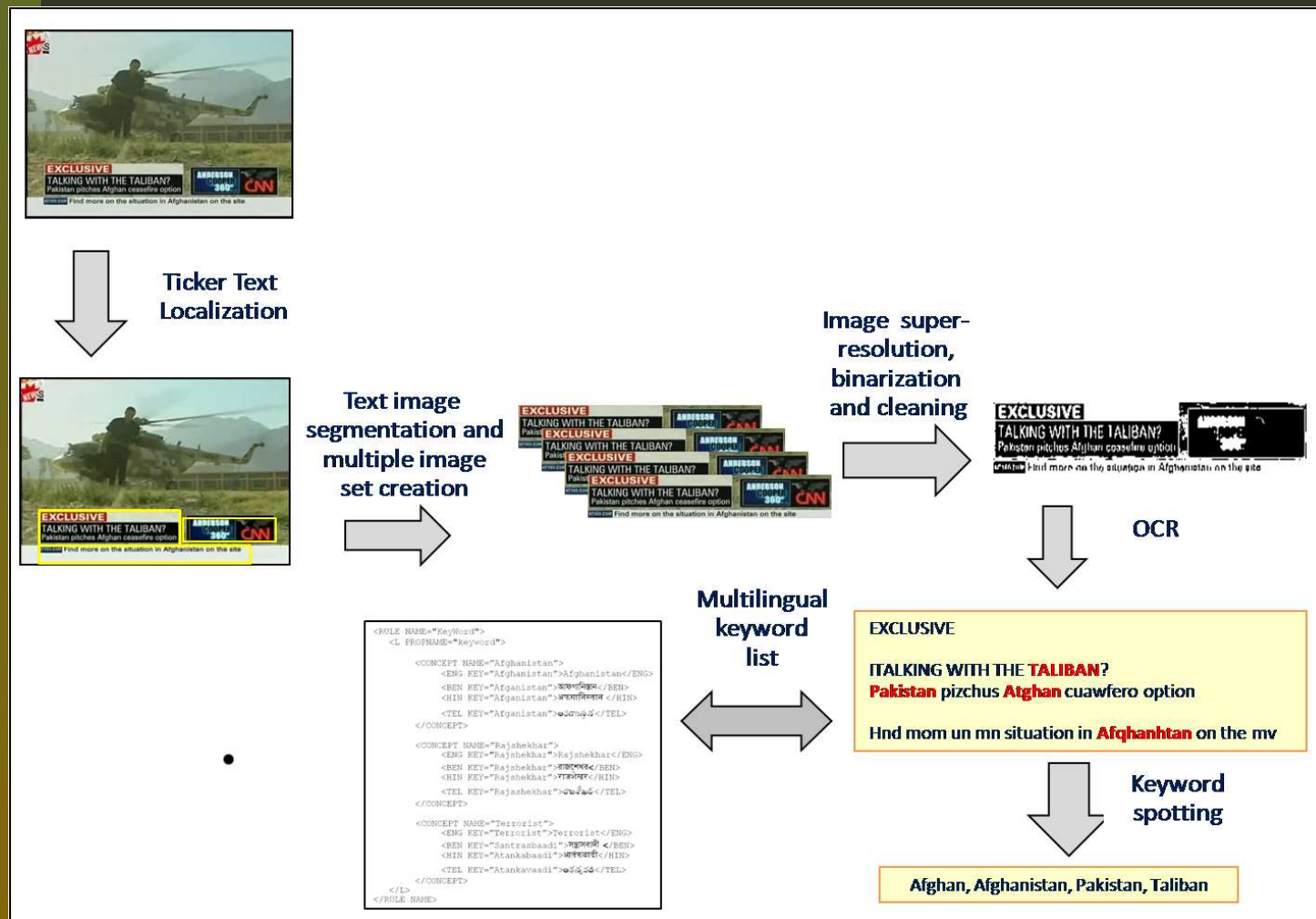
- Using keyword list in different languages
- English
 - Video OCR
 - sufficiently mature
 - increased accuracies with dynamic keyword list

How is this done?

Video OCR (English)



Video OCR (English)



Does it work for Indian Languages?

Indian Language Script OCR

Transcription	śivō rakṣatu gīrvāṇabhāṣārasāsvādatatparān
Bengālī	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতত্পরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸਿਵੇ ਰਕ੍ਸਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oriyā	ଶିବଃ । ରକ୍ଷତୁ ଗିର୍ବାଣଭାଷାରସାସ୍ବାଦତତ୍ପରାନ୍
Tamīl	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tēlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannada	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayālam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാന്
Grantha	ஸிவொ ரக்ஷதௌ ூர்வாணஹாஷாரஸாஸ்வாததத்பராந்

Source: http://www.myscribeweb.com/Phrase_sanskrit.png

- Script - Complex; Work being done in few languages
- Poor accuracies ;-(using approaches for English

Indian Language Script OCR

Transcription	śivō rakṣatu gīrvāṇabhāṣārasāsvādatatparān
Bengālī	শিবো রক্ষতু গীর্বাণভাষারসাস্বাদতত্পরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸਿਵੇ ਰਕ੍ਸਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oriyā	ଶିବଃ । ରକ୍ଷତୁ ଗିର୍ବାଣଭାଷାରସାସ୍ବାଦତତ୍ପରାନ୍
Tamil	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tēlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannada	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayālam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാന്
Grantha	ஸிவொ ரக்ஷதௌ ூர்வூணஹாஷாரஸாஸ்வாததத்பராந்

Source: http://www.myscribeweb.com/Phrase_sanskrit.png

- Script - Complex; Work being done in few languages
- Poor accuracies ;-(using approaches for English

New approach?

Recognition Free Approach? Maybe

- We know the keyword list (dynamic and update!)
- Generate images of keywords
(different fonts and sizes; usually not very different!)
- Match in the image space
 - ticker text can be segmented into word images;
 - compare with generated keyword images;
 - some work done at IIT Hyderabad
(<http://cvit.iiit.ac.in/projects/videoprocessing/>)
- Recognition accuracies still poor

Recognition Free Approach? Maybe

- We know the keyword list (dynamic and update!)
- Generate images of keywords
(different fonts and sizes; usually not very different!)
- Match in the image space
 - ticker text can be segmented into word images;
 - compare with generated keyword images;
 - some work done at IIT Hyderabad
(<http://cvit.iit.ac.in/projects/videoprocessing/>)
- Recognition accuracies still poor

*Recognition accuracies neither good in audio
nor in visual can we do something?*

Integrate

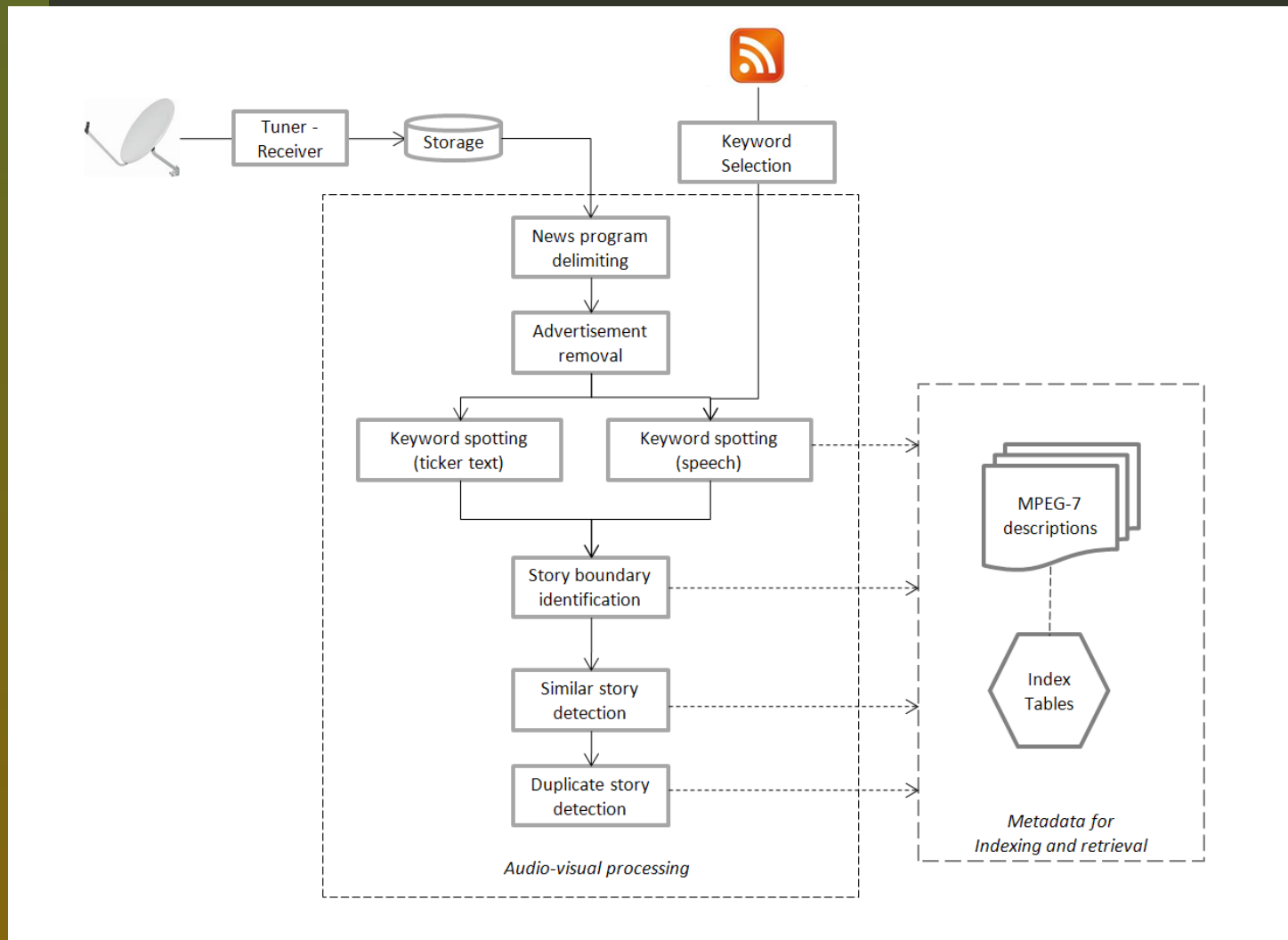
- Fuse the output of
 - audio channel keyword spotting and
 - keyword spotting in visual
- To get better keyword spotting accuracies in the news broadcast
 - in terms of precision
 - in terms of recall
- Validated experimentally (EUVIP 2010, IJDMB 2010)

Integrate

- Fuse the output of
 - audio channel keyword spotting and
 - keyword spotting in visual
- To get better keyword spotting accuracies in the news broadcast
 - in terms of precision
 - in terms of recall
- Validated experimentally (EUVIP 2010, IJDMB 2010)

In Summary ...

How is this used?



Approach Summary

- Use RSS feed to create keyword list
(in English; use NL processing)
- Identify English keyword in other Indian languages
(common nouns use word dictionary; transliteration for proper nouns)
- Create a multilingual keyword list
(source for keyword spotting)
- Keyword spotting in audio
(in different languages; same acoustic models; use Sphinx)
- Keyword spotting in visual
(ticker text in different languages; Recognition free)

What is Novel?

- Smart Use of RSS feed to construct a keyword list
- Results in dynamic and small KW list
increased keyword spotting accuracies
- Creation of a multilingual keyword list
- Using on-line resources
dictionary and transliteration
- Combining audio and visual to improve KW spotting
when acoustics data is small and the script to be recognized is complex!

What is Novel?

- Smart Use of RSS feed to construct a keyword list
- Results in dynamic and small KW list
increased keyword spotting accuracies
- Creation of a multilingual keyword list
- Using on-line resources
dictionary and transliteration
- Combining audio and visual to improve KW spotting
when acoustics data is small and the script to be recognized is complex!

Special (difficult?) problems need appropriate solutions ...

Results

- Multi lab effort
- Published
 - Indexing of Multilingual News Telecast using Audio Visual Keywords Second European Workshop on Visual Information Processing (EUVIP 2010).
 - Khare, "*Multimodal Indexing of Multilingual News Video*", International Journal of Digital Multimedia Broadcasting, 2010, Article ID 486487, 18 page.

Thank You

- Queries?
- Comments
- Suggestions?

SunilKumar.Kopparapu@TCS.Com
TCS Innovation Lab - Mumbai
Tata Consultancy Services Limited
Yantra Park, Thane (West), India.