Keyword Based Indexing of Multilingual News Videos

Sunil Kumar Kopparapu

SunilKumar.Kopparapu@TCS.COM

TCS Innovation Labs - Mumbai

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

December 2009

The 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

The 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 ...

The Scene

- Post Broadcast
  - 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

- Post Broadcast
  - 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 Scene

- Why not for Indian channels?
  - Simple. **No** closed captioned text. Unlike most Western channels!
  - No searchable text associated with news broadcast
- Who would need it anyway?
  - Security Agencies (get cues to monitor happenings)
  - Individuals (Searchable information on media sharing websites)
  - Broadcast channels (reuse the feed as per need)

The Scene

- Why not for Indian channels?
  - Simple. **No** closed captioned text. Unlike most Western channels!
  - No searchable text associated with news broadcast
- Who would need it anyway?
  - Security Agencies (get cues to monitor happenings)
  - Individuals (Searchable information on media sharing websites)
  - 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
  - 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
    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
  - 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
    absence of closed caption text

How can we do this?

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
- 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
- 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

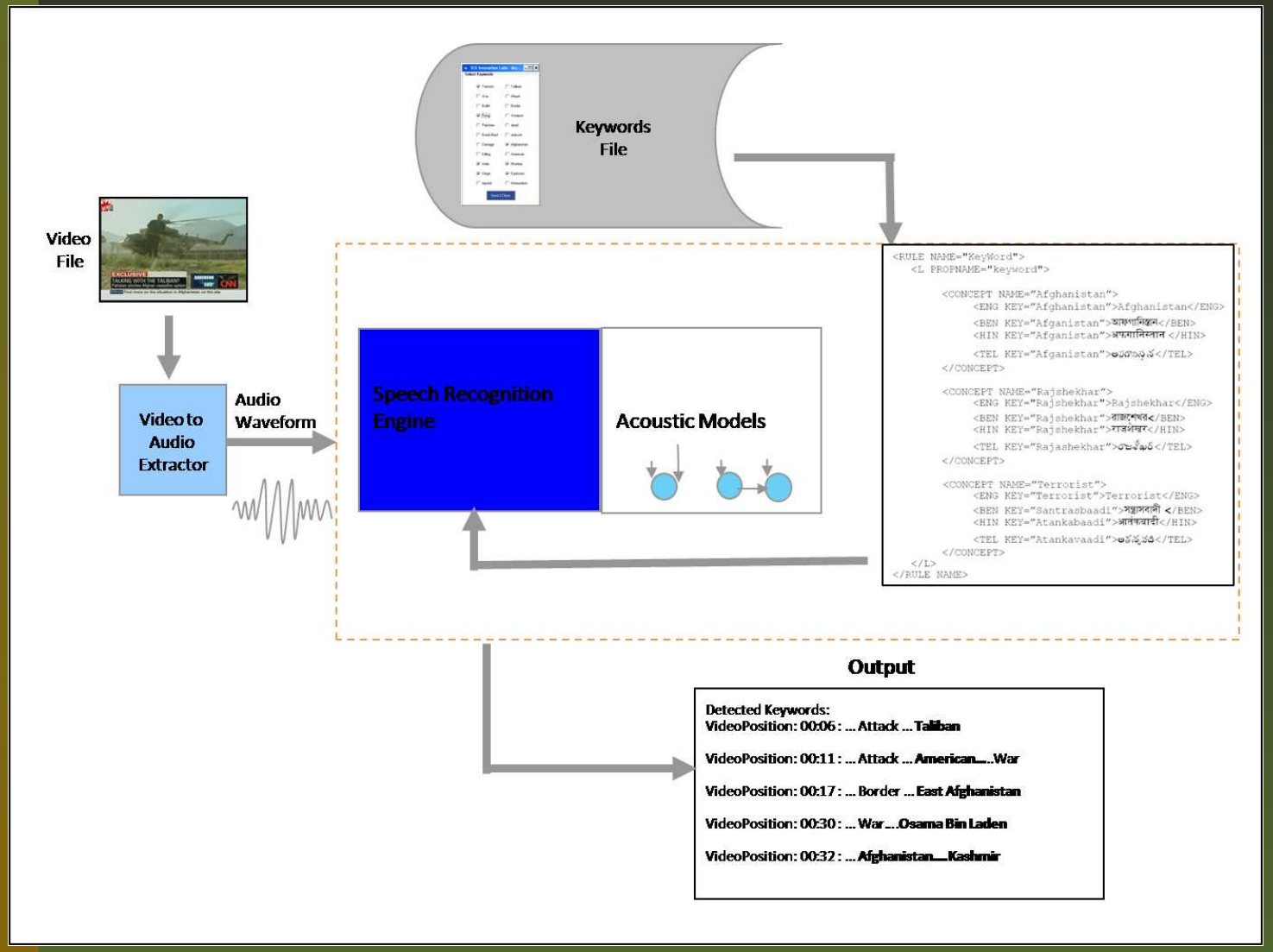
```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

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

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">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>
```

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
  (if the keywords are words not in the dictionary - we add the pronunciations)
- Works equally well with Hindi!
  (pronunciations lexicon for Hindi 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
  (if the keywords are words not in the dictionary - we add the pronunciations)
- Works equally well with Hindi!
  (pronunciations lexicon for Hindi keywords words also added)

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

Keyword spotting in visual

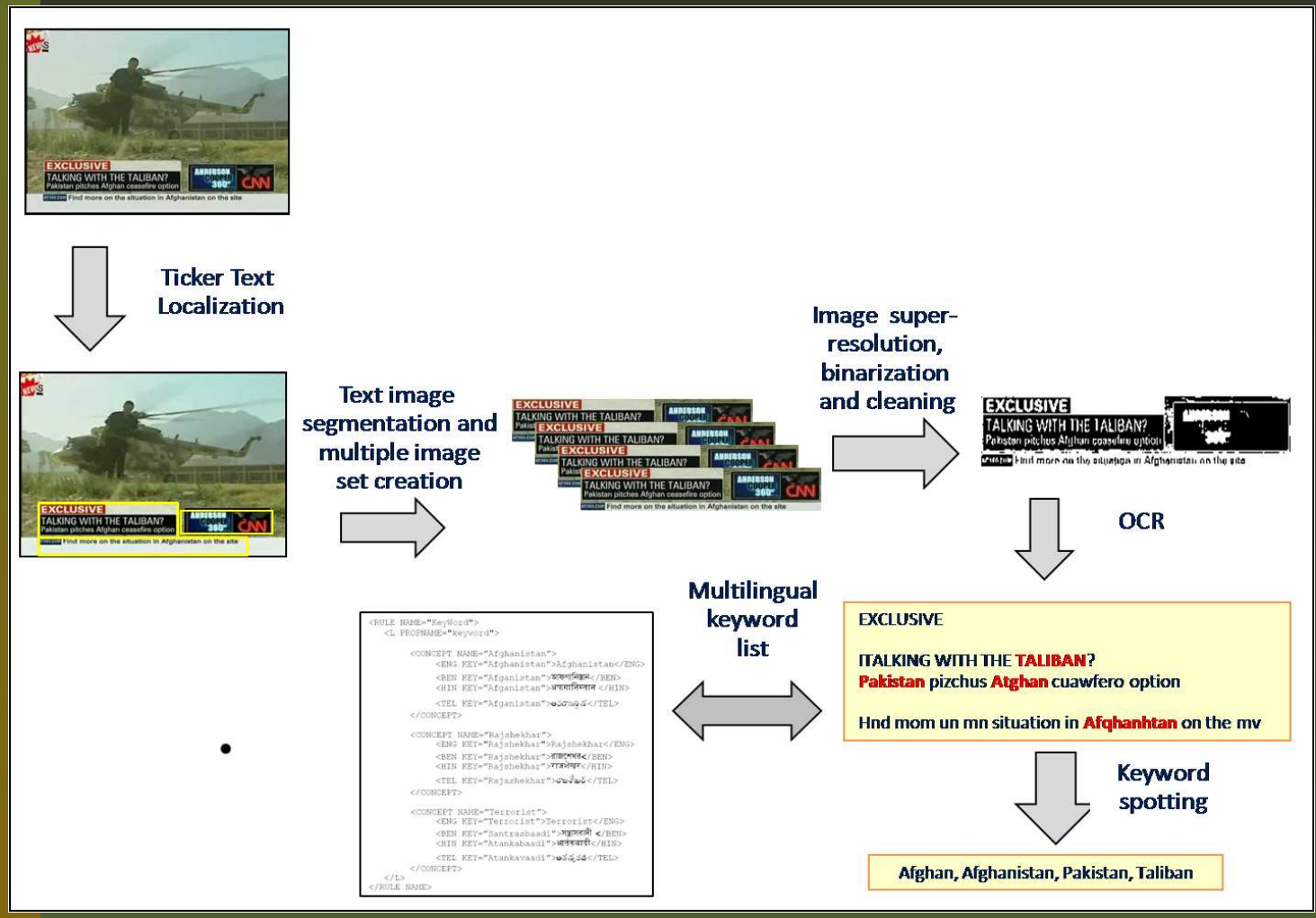
- Using keyword list in different languages
- English
  - Video OCR
  - sufficiently mature
  - increased accuracies with dynamic keyword list

Keyword spotting in visual

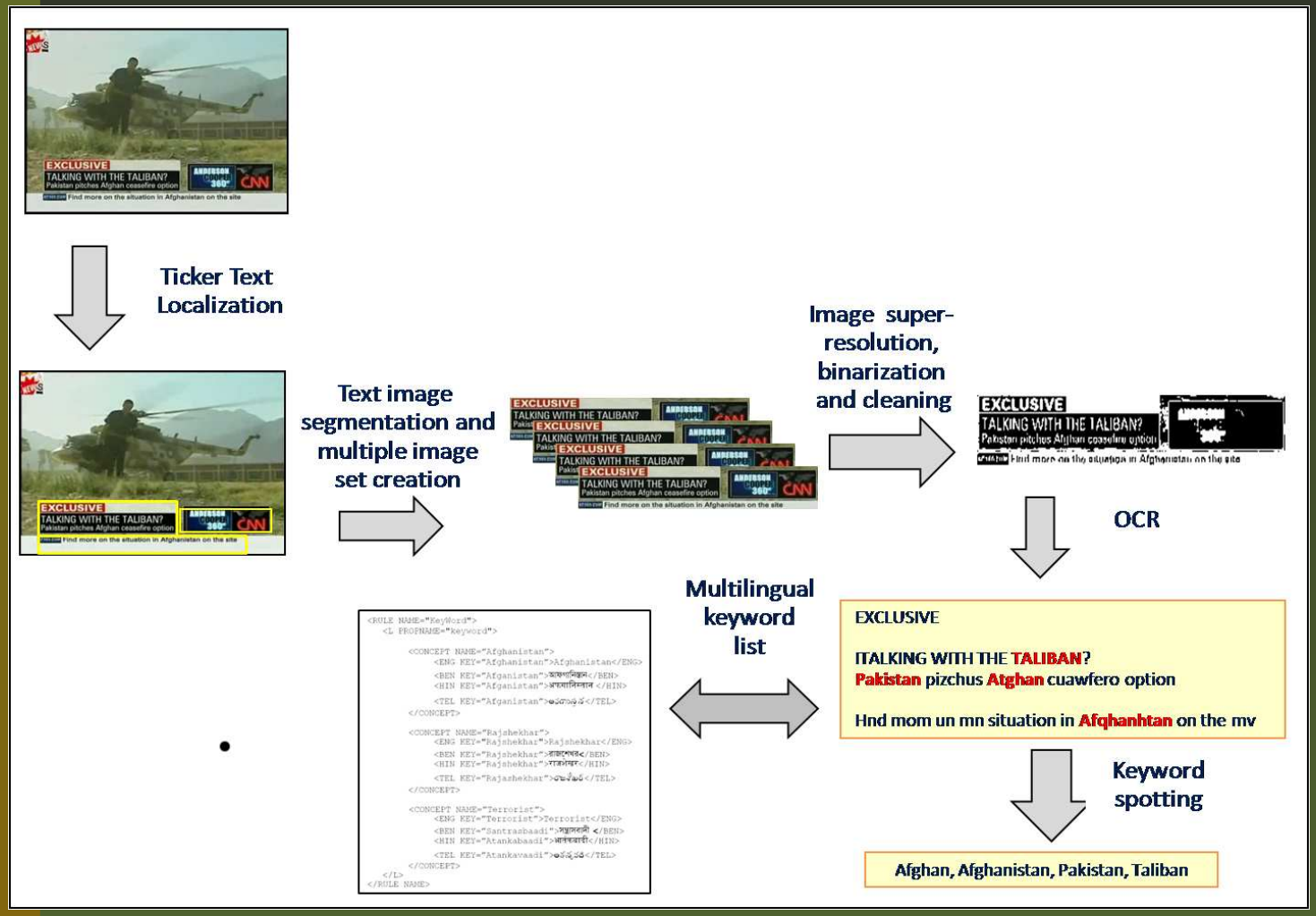
- Using keyword list in different languages
- English
  - Video OCR
  - sufficiently mature
  - increased accuracies with dynamic keyword list

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 গীর্বাণভাষারসাস্বাদতৎপরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸ਼ਿਵੋ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oṛiyā	ଶିବୋ ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ୱାଦତତ୍ପରାନ୍
Tamiḻ	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tĕlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannaḍa	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayāḷam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാൻ
Grantha	𑌶𑌿𑌵𑍋 𑌰𑌕𑍍𑌷𑌤𑍁 𑌗𑍀𑌰𑍍𑌵𑌾𑌣𑌭𑌾𑌷𑌾𑌰𑌸𑌾𑌸𑍍𑌵𑌾𑌦𑌤𑌤𑍍𑌪𑌰𑌾𑌨𑍍

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 গীর্বাণভাষারসাস্বাদতৎপরান্
Devanāgarī	शिवो रक्षतु गीर्वाणभाषारसास्वादतत्परान्
Gujarātī	શિવો રક્ષતુ ગીર્વાણભાષારસાસ્વાદતત્પરાન્
Gurmukhī	ਸ਼ਿਵੋ ਰਕ੍ਸ਼ਤੁ ਗੀਰ੍ਵਾਣਭਾਸ਼ਾਰਸਾਸ੍ਵਾਦਤਤ੍ਪਰਾਨ੍
Oṛiyā	ଶିବୋ ରକ୍ଷତୁ ଗୀର୍ବାଣଭାଷାରସାସ୍ୱାଦତତ୍ପରାନ୍
Tamiḻ	ஷிவோ ரக்ஷது கீர்வாணபாஷாரஸாஸ்வாததத்பராந்
Tĕlugu	శివో రక్షతు గీర్వాణభాషారసాస్వాదతత్పరాన్
Kannaḍa	ಶಿವೋ ರಕ್ಷತು ಗೀರ್ವಾಣಭಾಷಾರಸಾಸ್ವಾದತತ್ಪರಾನ್
Malayāḷam	ശിവോ രക്ഷതു ഗീർവാണഭാഷാരസാസ്വാദതത്പരാൻ
Grantha	𑌶𑌿𑌵𑍋 𑌰𑌕𑍍𑌷𑌤𑍁 𑌗𑍀𑌰𑍍𑌵𑌾𑌣𑌭𑌾𑌷𑌾𑌰𑌸𑌾𑌸𑍍𑌵𑌾𑌦𑌤𑌤𑍍𑌪𑌰𑌾𑌨𑍍

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 images space
  - ticker text can be segmented into word images;
  - compare with generated keyword images;
  - some work done at IIIT 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 images space
  - ticker text can be segmented into word images;
  - compare with generated keyword images;
  - some work done at IIIT Hyderabad
    (http://cvit.iiit.ac.in/projects/videoprocessing/)
- Recognition accuracies still poor

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

Integrate

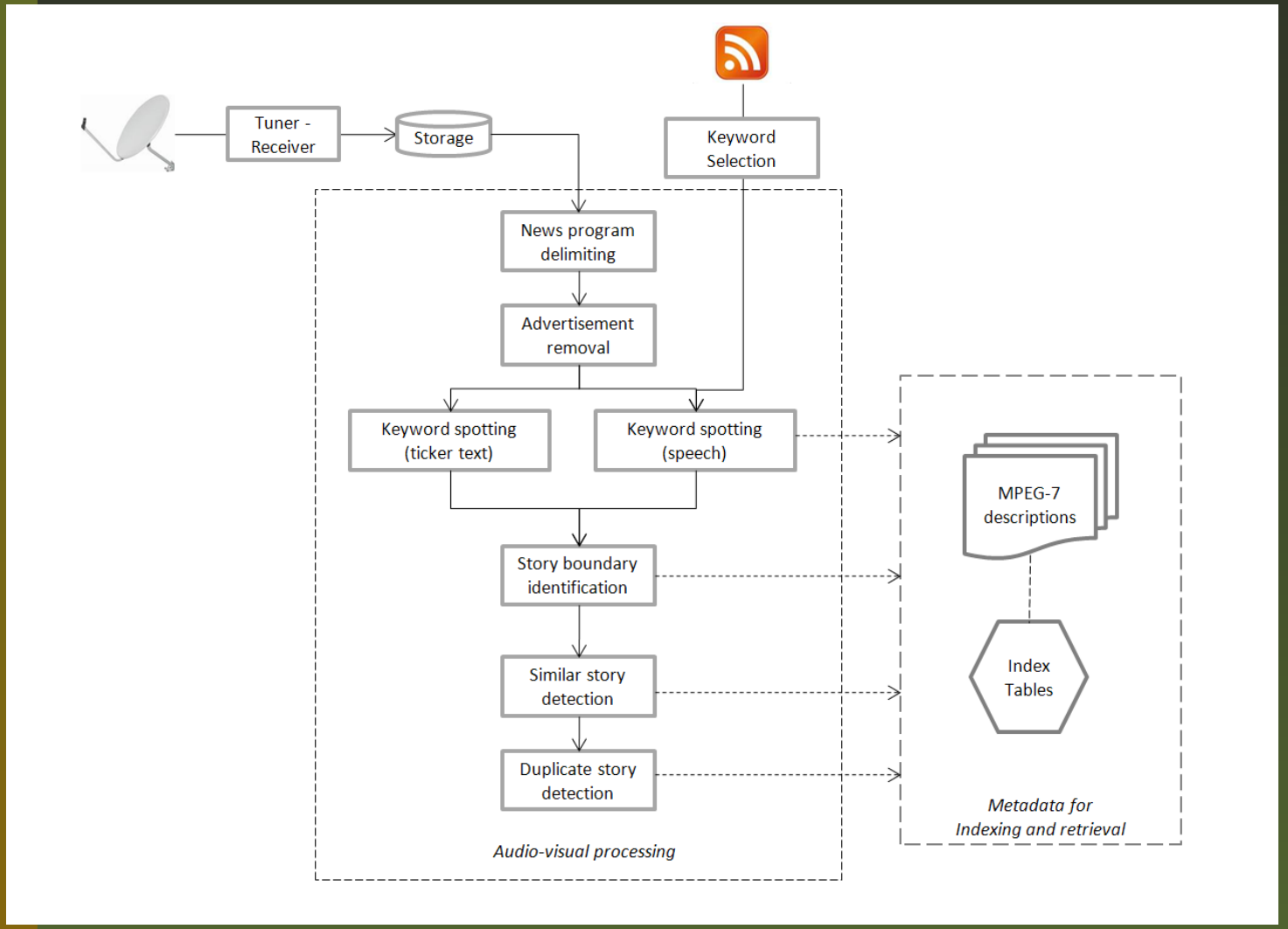
- Fuse the output of
  - audio channel keyword spotting and
  - keyword spotting in visual
- To get better keyword spotting accuracies in the news broadcast
  - in terms of precision
  - in terms of recall
- Validated by some preliminary experiments

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 by some preliminary experiments

In Summary ...

Summary


Tuner - Receiver
Storage
Keyword Selection
News program delimiting
Advertisement removal
Keyword spotting (ticker text)
Keyword spotting (speech)
Story boundary identification
Similar story detection
Duplicate story detection
Audio-visual processing
MPEG-7 descriptions
Index Tables
Metadata for Indexing and retrieval

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 ...

Thank You

- Queries?
- Comments
- Suggestions?

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