Keyword Based Indexing of Multilingual News Videos

Sunil Kopparapu and Hiranmay Ghosh

SunilKumar.Kopparapu@TCS.COM

TCS Innovation Labs - Mumbai

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

December 2009

Summary

- The Scene (Indian News broadcasting)
  - Television broadcast in 10 **languages**
  - **No** closed captioned text
  - **Need** for annotating multilingual videos **exists**
  - Indexing based **only** on audio and visual cues
- Status
  - **Much desired** in audio and visual processing for Indian languages

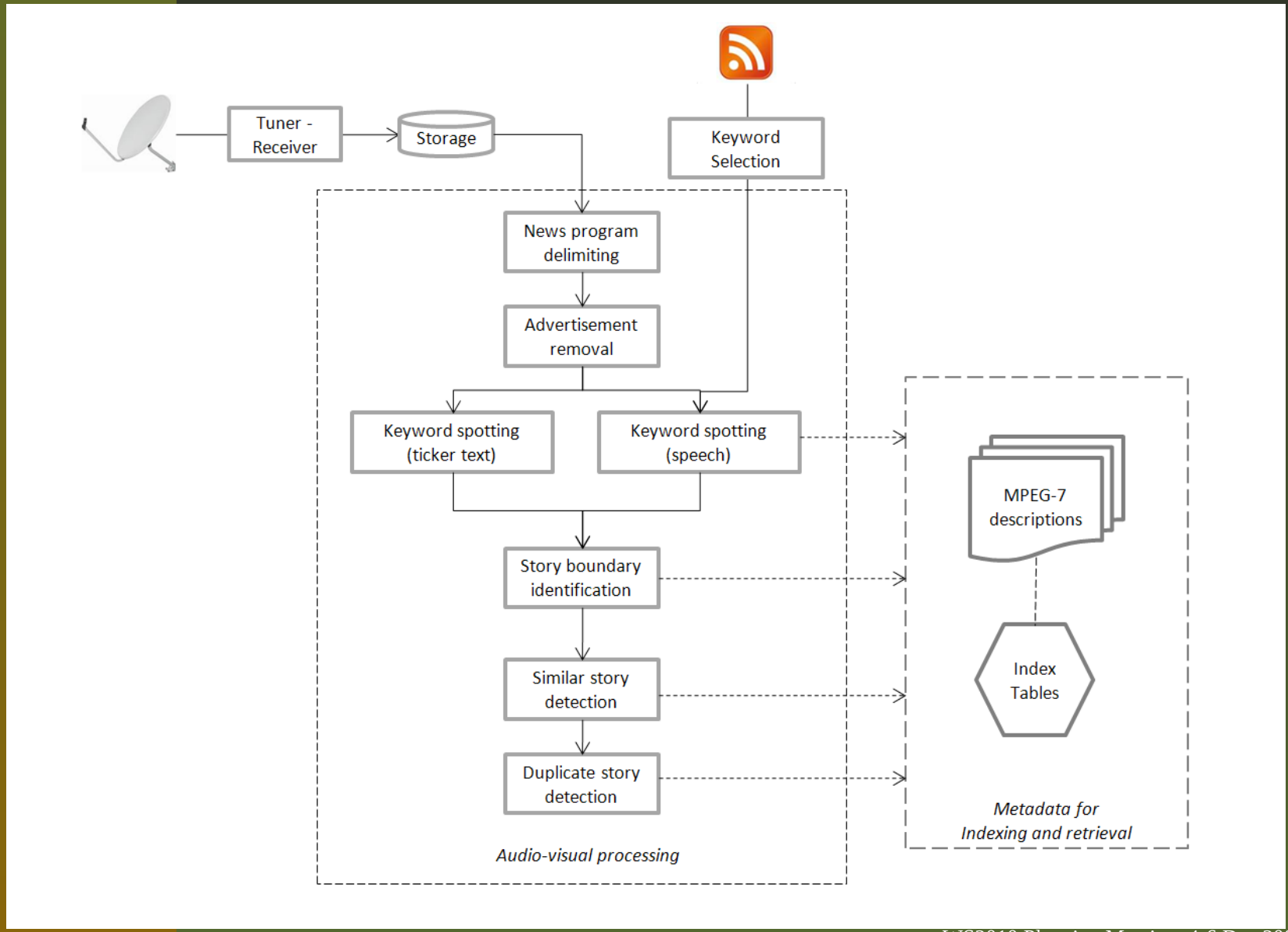
Summary

- The Scene (Indian News broadcasting)
  - Television broadcast in 10 **languages**
  - **No** closed captioned text
  - **Need** for annotating multilingual videos **exists**
  - Indexing based **only** on audio and visual cues
- Status
  - **Much desired** in audio and visual processing for Indian languages

In this scenario can we do something to enable indexing of multilingual videos? Using orthogonal cues? Assist in cross-lingual search

....

Approach Overview



Dynamic Keyword list

- RSS feed from Internet (several exist)
- Use RSS feed to create keyword list (in English). (Using some NL tools (Named entity, statistical n-gram analysis, ..))
- Keywords in news broadcast are proper nouns and common nouns
- Identify English keyword equivalents in other Indian languages (How?)
  (use Technology Development for Indian Language (TDIL) translation tools .. or word dictionaries)

Sample Multilingual Keyword list

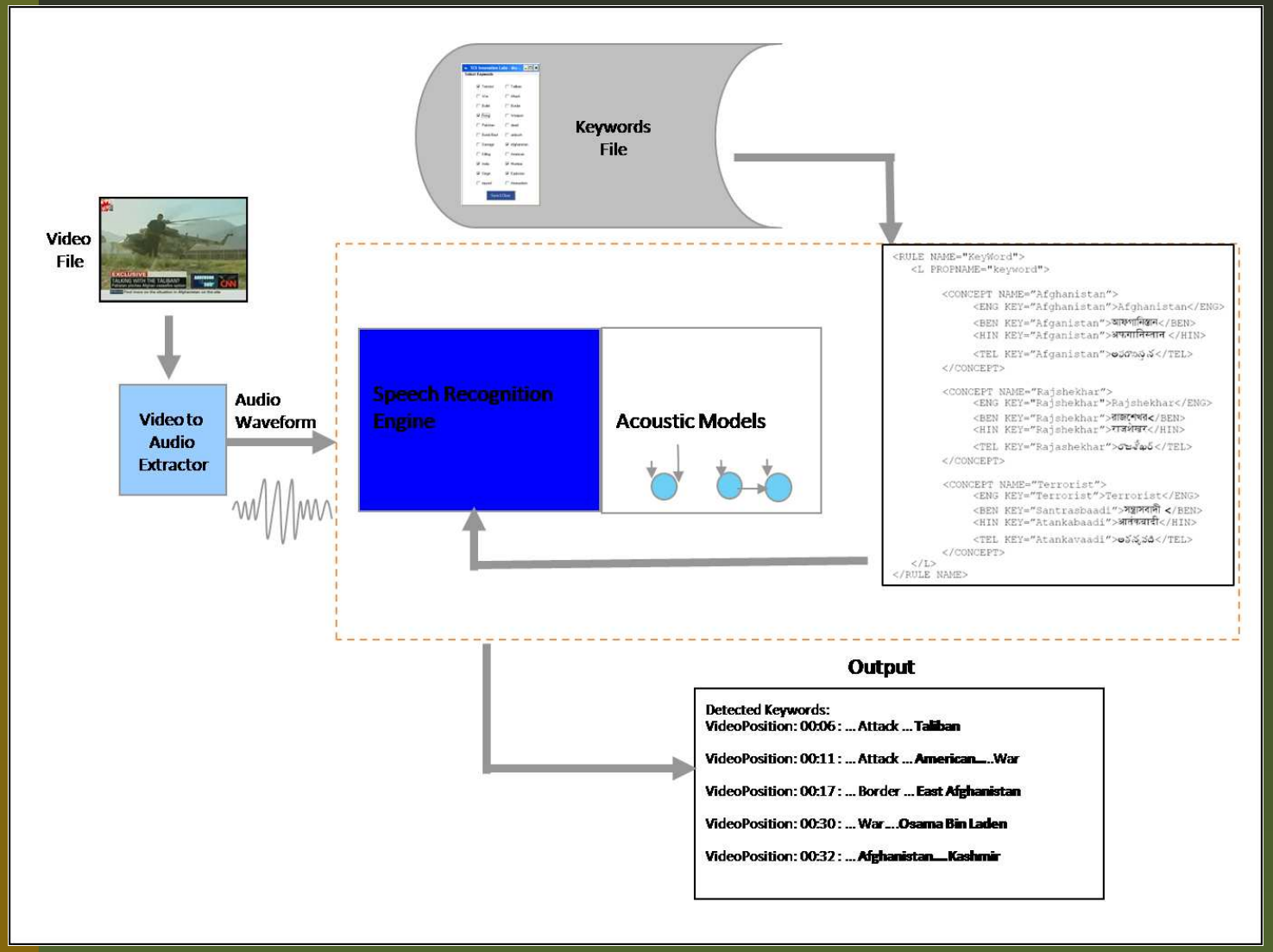
```
<RULE NAME="KeyWord">
   <L PROPNAME="keyword">

        <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>
```

Audio KW Spotting



Challenges in Audio KWS

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

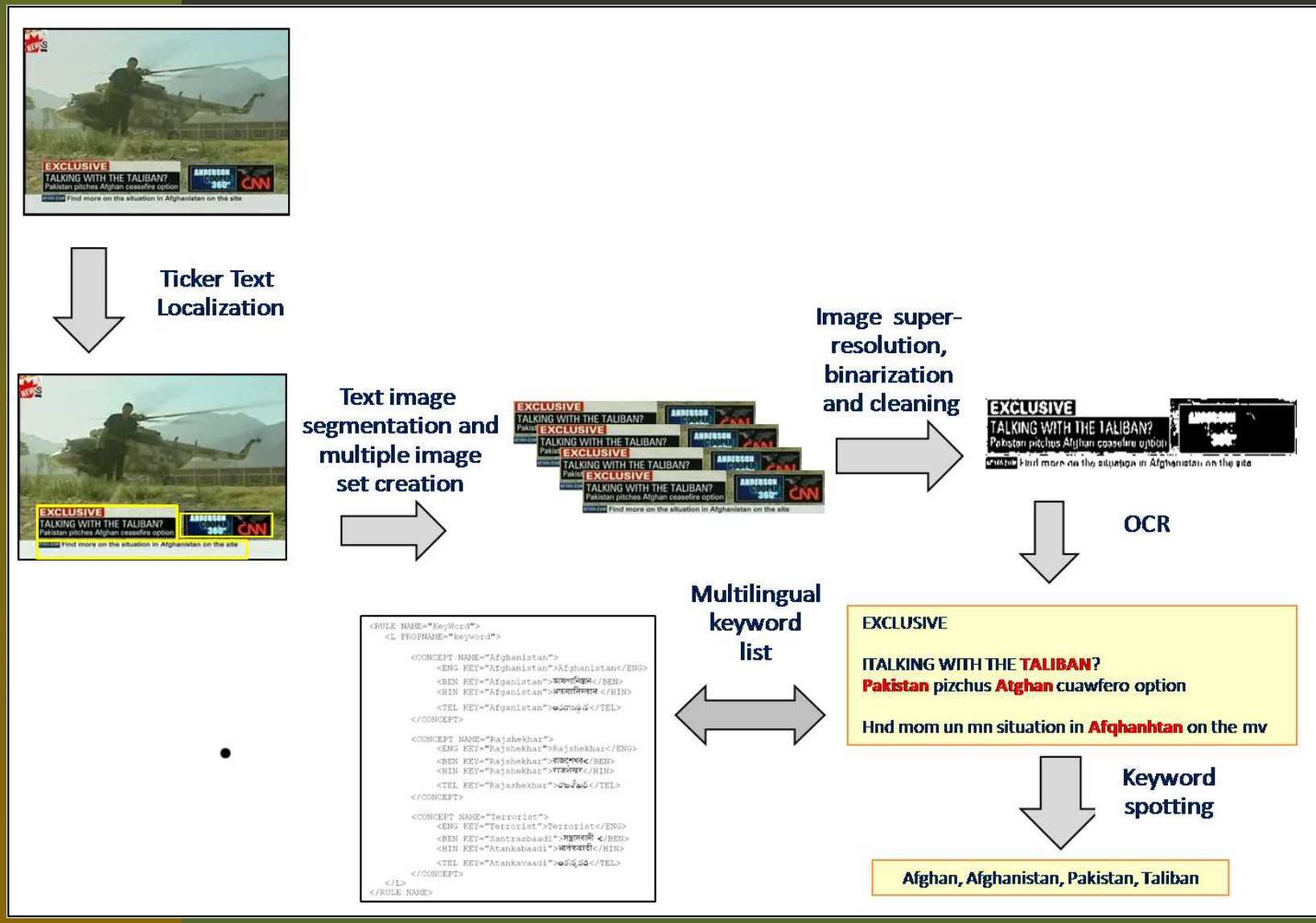
  Can we do something?
- Build or use acoustics models for one Indian language
  (simpler than building for all! Can we use English acoustic models??)
- Use this for keyword spotting
  (Largely Indian language phonetic and we are doing only keyword spotting anyway!)

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!

Video OCR (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

- Indian Language Script - Complex
- Work being done in few languages

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/))

Approach

- 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 accuracies)
- Creation of a multilingual keyword list
- Using on-line resources; dictionary and transliteration
- Combining audio and Video OCR to improve KW spotting (when acoustics data is small and the script to be recognized is complex!)

In six weeks?

- Dec 09 - May 10 (Data source identified or collected; Framework and tools shortlisted, sample multilingual keyword list created)
- Wk1: Keywords from RSS feed || Video OCR English || Audio KWS English
- Wk2: Creation of multilingual KW list || Video OCR (L_1)|| Pronunciation Lexicon
- Wk3: Video OCR (L_1)|| Audio KWS (L_1 and L_2); same acoustic models
- Wk4: Video OCR (L_2) || Audio KWS (L_1 and L_2)
- Wk5:Integration (Video OCR and audio KWS)
- Wk6:Testing || Indexing || Report

Thank You

- Should give a good platform to mix and match expertise in the areas of NL, Script and Speech
- Scope small? Only Indian languages? May be ideas will emerge that can be used else where!

Thank You

- Should give a good platform to mix and match expertise in the areas of NL, Script and Speech
- Scope small? Only Indian languages? May be ideas will emerge that can be used else where!

SunilKumar.Kopparapu@TCS.Com
TCS Innovation Lab - Mumbai
Tata Consultancy Services Limited

Yantra Park, Thane (West), India.