MILE lab
Electrical Engineering
Indian Institute of Science (IISc)
Bangalore

MAST

Multi-Script Annotation Toolkit for Scenic Text

T Kasar
Deepak Kumar
M N Anil Prasad
D Girish
Prof. A G Ramakrishnan

Motivation

- Need for Large annotated data-base
  - Training & bench marking
- Camera-based documents
  - only document image annotation
- Efficient annotation of Scenic text
  - Multi script : flexible adaptation
  - Pixel level ground truth
  - Open source

MILE_MAST_GT

MAST: Multi-script Annotation toolkit for Scenic Text

WORD LEVEL

LOAD

SEGMENT WORD

THRESHOLD 25

RESEGMENT

RELOAD

BINARIZE

SELECT SCRIPT

NEXT WORD

SAVE

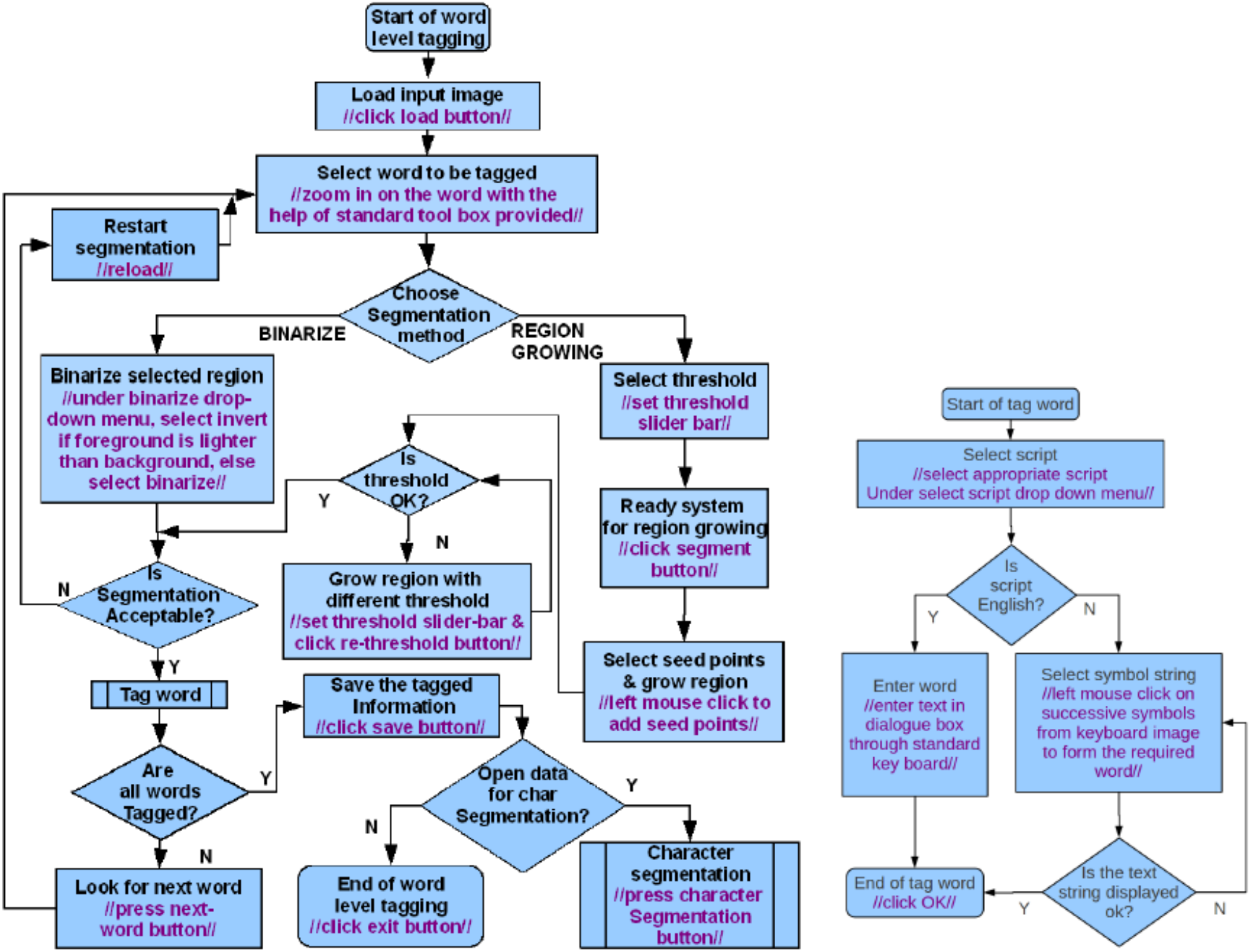
WORD BOUNDARY

SEGMENT CHAR

EXIT

Methodology

- Region growing around manual seed selection
  - CIE L*a*b* color space, with Euclidean distance measure
  - Vary threshold such that pixels corresponding to text are extracted
- If required use Otsu's binarization
- Multiple round :tagging a word at a time


Start of word level tagging
Load input image //click load button//
Select word to be tagged //zoom in on the word with the help of standard tool box provided//
Restart segmentation //reload//
Choose Segmentation method
BINARIZE
REGION GROWING
Binarize selected region //under binarize drop-down menu, select invert if foreground is lighter than background, else select binarize//
Select threshold //set threshold slider bar//
Is threshold OK?
Y
N
Grow region with different threshold //set threshold slider-bar & click re-threshold button//
Ready system for region growing //click segment button//
Select seed points & grow region //left mouse click to add seed points//
Is Segmentation Acceptable?
N
Y
Tag word
Save the tagged Information //click save button//
Are all words Tagged?
Y
N
Look for next word //press next-word button//
Open data for char Segmentation?
Y
N
End of word level tagging //click exit button//
Character segmentation //press character Segmentation button//
Start of tag word
Select script //select appropriate script Under select script drop down menu//
Is script English?
Y
N
Enter word //enter text in dialogue box through standard key board//
Select symbol string //left mouse click on successive symbols from keyboard image to form the required word//
Is the text string displayed ok?
Y
N
End of tag word //click OK//

MILE_KANNADA_UNICODE_INPUT

ಆ	ರ	್ಪೋ	ಗ	್	ಯ

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ	ಂ	ಃ
್	ಾ	ಿ	ೀ	ು	ೂ	ೃ	ೄ	ೆ	ೇ	ೈ	ೊ	ೋ	ೌ	---	---
ಕ	ಖ	ಗ	ಘ	ಙ	ಚ	ಛ	ಜ	ಝ	ಞ	ಟ	ಠ	ಡ	ಢ	ಣ	---
ತ	ಥ	ದ	ಧ	ನ	ಪ	ಫ	ಬ	ಭ	ಮ	---	---	---	---	ಫ಼	ऽ
ಯ	ರ	ಱ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ	ೞ	---	---	ೱ	ೲ	ಁ
೦	೧	೨	೩	೪	೫	೬	೭	೮	೯	---	---	---	---	ॅ	ಀ

OK

UNDO

CANCEL

Ground truth at word level

- Img.jpg

- Img_seg.jpg
- Img.txt

Bounding box

Script

Word

```
DSC02617_1.jpg 459 386 195 724 KANNADA ಆರೋಗ್ಯ
DSC02617_2.jpg 453 1167 169 497 KANNADA ಕೇಂದ್ರ
DSC02617_3.jpg 826 567 127 560 DEVANAGARI स्वास्थ्य
DSC02617_4.jpg 762 1195 200 334 DEVANAGARI केन्द्र
DSC02617_5.jpg 1136 276 127 731 ENGLISH HEALTH
DSC02617_6.jpg 1127 1080 125 739 ENGLISH CENTRE
```

- IMG
  - Img_1.jpg
  - .
  - .
  - .

mile
MEDICAL INTELLIGENCE AND
LANGUAGE ENGINEERING LAB

Issues in Character level annotation

- A Connected component (CC) need not correspond to a character!!!

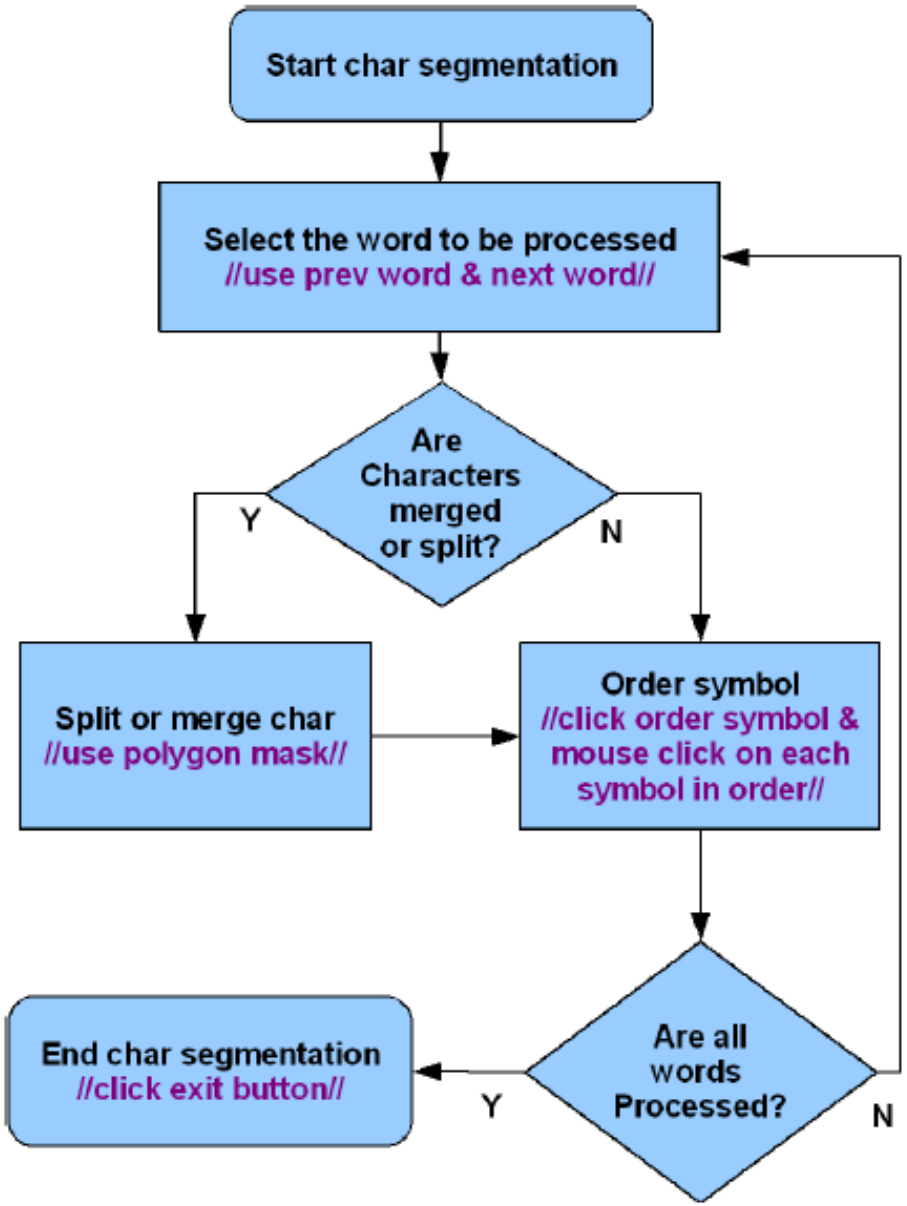
  - In English ‘i’
  - Many Indic languages, a word is a CC

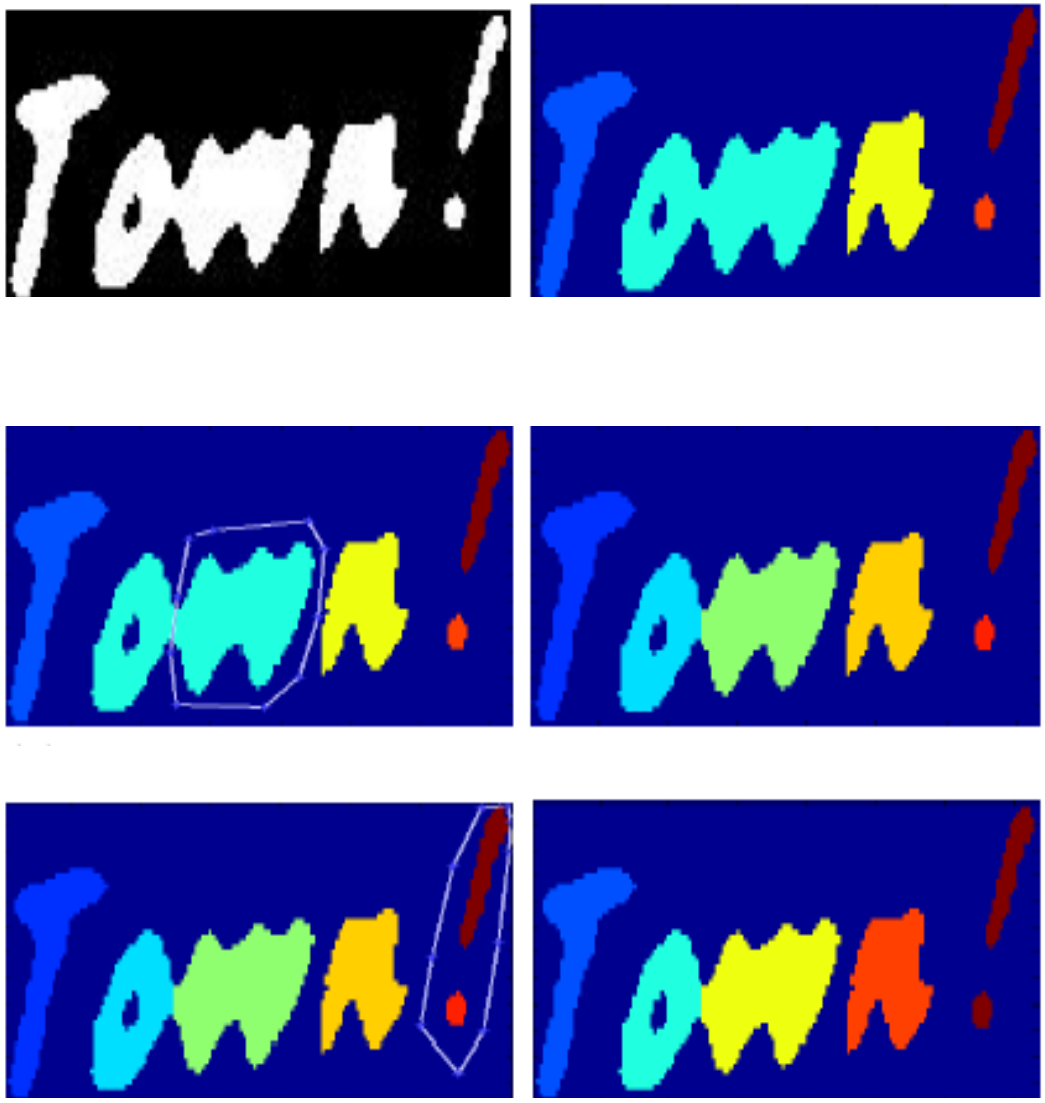



- Therefore need to split or merge CC
- Symbol level as required by the user

mile
MEDICAL INTELLIGENCE AND
LANGUAGE ENGINEERING LAB

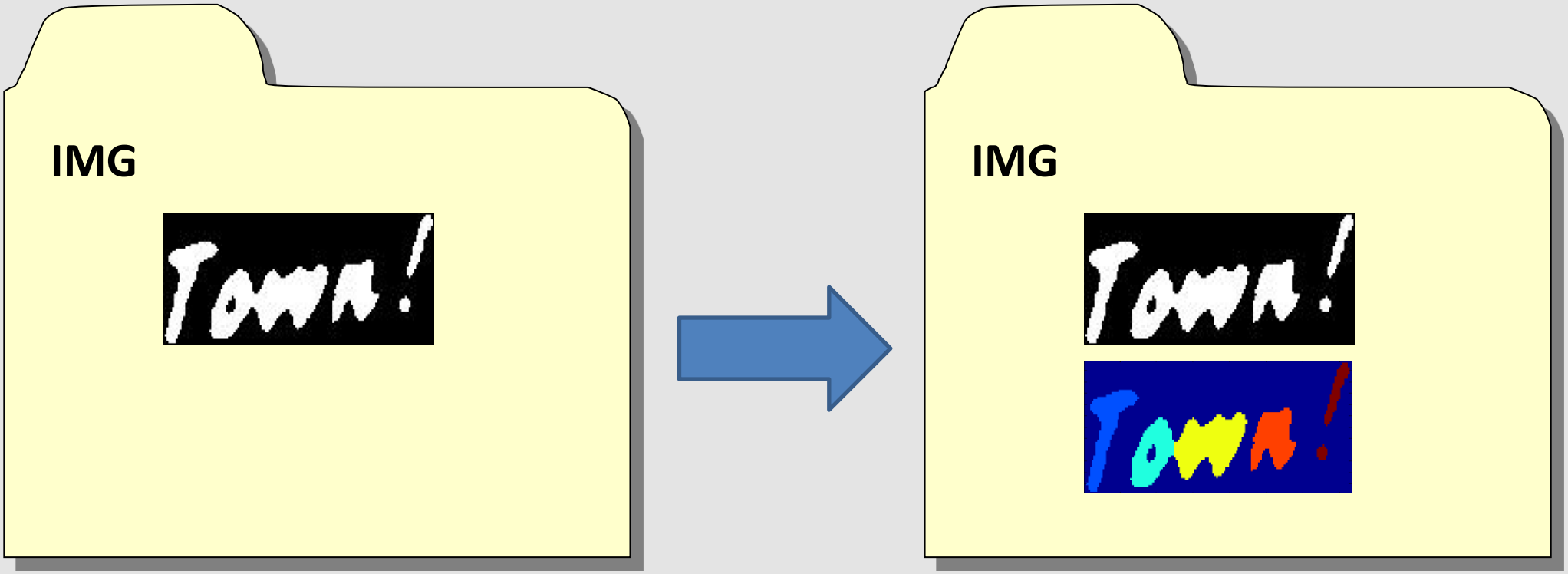
Char level





mile
MEDICAL INTELLIGENCE AND
LANGUAGE ENGINEERING LAB

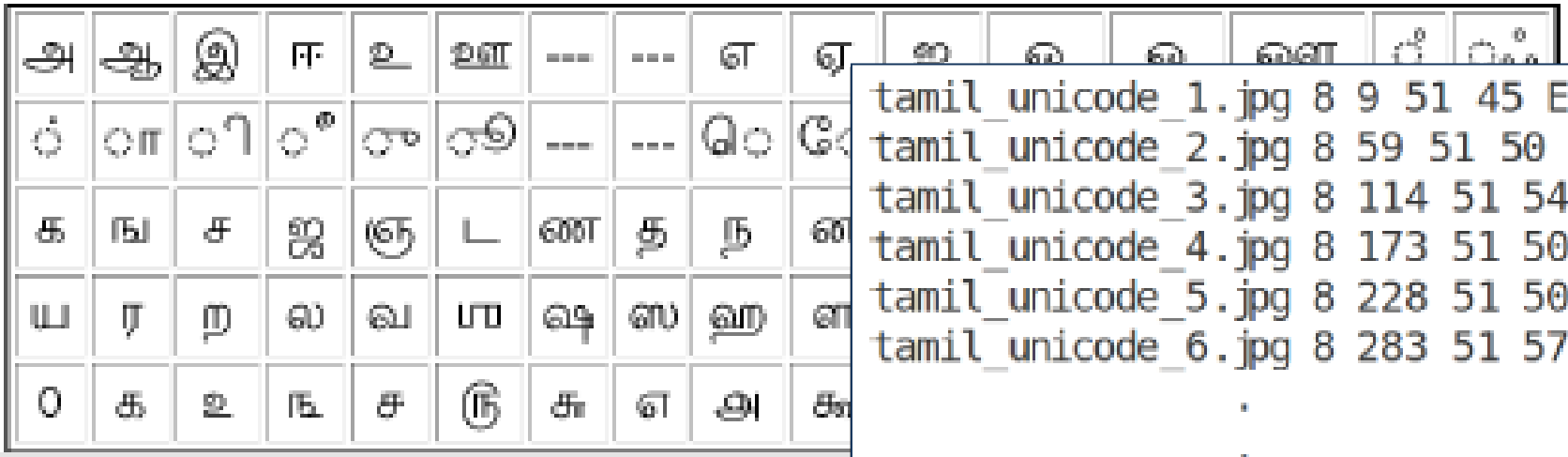
Ground truth at symbol level



mile
MEDICAL INTELLIGENCE AND
LANGUAGE ENGINEERING LAB

Generation of new keyboard

- Use Keyboard image as i/p to word tagging module
- Perform region growing such that bounding box in the generated text file represents the button
- Tag the buttons Unicode using the standard English keyboard



```
tamil_unicode_1.jpg 8 9 51 45 ENGLISH b85
tamil_unicode_2.jpg 8 59 51 50 ENGLISH b86
tamil_unicode_3.jpg 8 114 51 54 ENGLISH b87
tamil_unicode_4.jpg 8 173 51 50 ENGLISH b88
tamil_unicode_5.jpg 8 228 51 50 ENGLISH b89
tamil_unicode_6.jpg 8 283 51 57 ENGLISH b8a
                    .
                    .
                    .
tamil_unicode_61.jpg 234 511 51 56 ENGLISH bf0
tamil_unicode_62.jpg 234 572 51 63 ENGLISH bf1
```

Downloadable Resources

- http://mile.ee.iisc.ernet.in/mast/
- MatLab code
- Annotated Database (will be available shortly)
- User guide

Future work

- Support for keyboard generation
- Multiple segmentation methodologies
- Semi-automation in tagging text
- Web based