

MAST: Multi-Script Annotation Toolkit for Scenic Text

T Kasar, D Kumar, M N Anil Prasad, D Girish and A G Ramakrishnan
Medical Intelligence and Language Engineering Laboratory
Indian Institute of Science, Bangalore, INDIA - 560 012
{tkasar, deepak, anilprasadm, dasarigirish, ramkiag}@ee.iisc.ernet.in

Abstract

There is a significant need for methods to extract and recognize text in scenes. Unlike the case of processing conventional document images, natural scene text understanding usually involves a pre-processing step of text region location and extraction before subjecting the acquired image for character recognition task. There are no standard, pixel-level annotated databases containing camera-captured multi-script, multi-oriented text. The availability of annotated datasets for scenic images will aid in testing and quantifying the performances of various document analysis and recognition algorithms. We have developed a semi-automatic tool to aid the creation of such annotated databases for research in camera-based document analysis. The procedure involves manual seed selection followed by a region growing process to segment each word present in the image. The threshold for region growing can be varied by the user so as to ensure pixel-accurate character segmentation. The text present in the image is tagged word-by-word. A virtual keyboard interface has also been designed for entering the ground truth in ten Indic scripts, besides English. The keyboard interface can easily be generated for any script, thereby expanding the scope of the toolkit. Optionally, each segmented word can further be labeled into its constituent characters/symbols. Polygonal masks are used to split or merge the segmented words into valid characters/symbols. The ground truth is represented by a pixel-level segmented image and a '.txt' file that contains information about the number of words in the image, word bounding boxes, script and ground truth Unicode. The toolkit, which we call MAST, can be used to generate ground truth and annotation for generic document images. The software, developed on Matlab, is available online¹ along with a detailed description of the functionalities of each of the menu items. We hope that researchers worldwide will find it useful in creating ground truth database for any generic document image and performance evaluation.

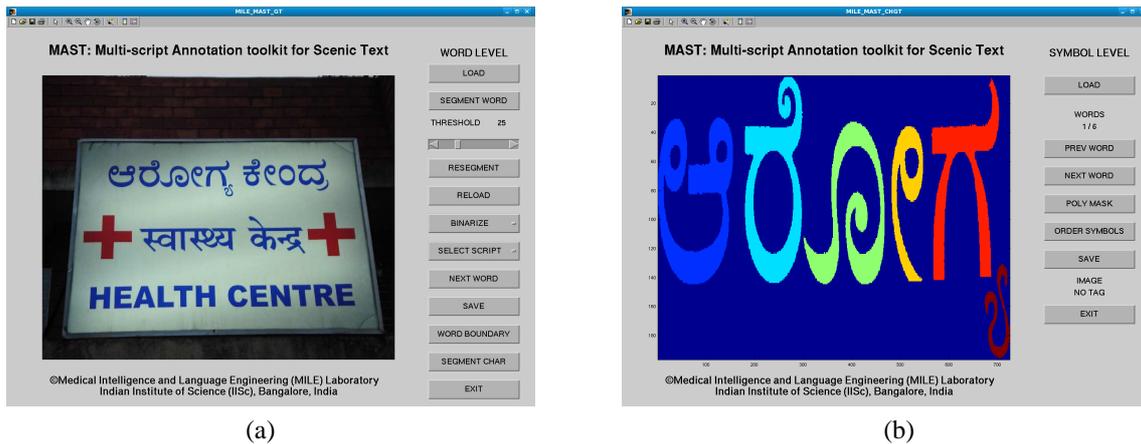


Figure 1. Screenshot of the user-interface for (a) word-level and (b) symbol level annotation.

¹<http://mile.ee.iisc.ernet.in/mast>