Abstract

Images are important information carriers which are often used in email messages and web pages to attach textual information. In Born digital compound image (BDCI) text and graphics/pictures come together on digital devices having certain distinct characteristics like low resolution (easy for online transmission and to display on screen) and text is created digitally on image. Text from BDCI can be effectively adopted for large numbers of applications like to retrieve contents of web, to improve indexing, to enhance content accessibility and content filtering. There are several problems to distinguish texts from BDCI because, text appears in various styles (i.e. Orientation, size, and colour), some neighbour texts are connected, and some text characters are superimposed on pictorial region which may lead to misclassification. Although researchers have proposed many methods in which character-level and block-based objects are commonly assumed to separate text from compound images. But these methods
failed to extract reliable features to detect all texts as well as to identify connected components. To address these issues, novel efficient algorithm Local Image Activity Measure (LIAM) and Scale and Orientation Invariant Grouping (SOIG) are proposed to assemble separated characters into Textual Connected Component (TCC). These algorithms are based on distribution of pixel variations and mean intrastring distance to precisely segment textual regions from BDCI.

References


**Index Terms**

Computer Science  

Image Processing

**Keywords**

Born-digital Compound Image  
Text Segmentation  
Mean Intrastring Distance.