Feature Extraction from Old Tamil Newspapers Using Histogram Minima

Shanmugalingam Kasthuri [1], Mahendrarajah Darsha, L.Ranathunga

Department of Information Technology, University of Moratuwa, Moratuwa, Sri Lanka

Archaeological records which provide information about the history of human cultures and past events. Newspapers can be considered as one of the main sources of gathering archaeological data. It can be seen that there exist only a few numbers of systems for the processing of old Tamil newspaper articles. An automated image processing system proposed as a suitable solution to the way of efficient and flexible searching approach, which can be used for old Tamil newspapers. In this paper is presented image processing technique to extract the features such as headlines and sub-headlines from old Tamil newspaper scanned images. Historical newspapers become damaged over time. The images of these newspapers become difficult to read the contents. The quality of the image improved by preprocessing techniques such as grayscale dilation, median filtering, and adaptive binarization. It helps to easily extract needed information on the image. Segment the article and identify the heading of the article will help to improve data manipulation. Feature extraction from old Tamil newspaper images followed these step processes; Horizontal smoothing is necessary to distinguish the paragraphs and empty space between each column; Vertical smoothing is implemented to distinguish between each paragraph and headlines; Logical AND operation combines the outcome of horizontal smoothing and vertical smoothing using AND operation; Height measurement of each block is followed by horizontal projection, that involves scanning of pixels through horizontal arrays to measure the black pixel density against index of each row by using horizontal histogram minima. This step identified horizontals breaking points of individual regions within an article. The four major horizontal regions are headlines, sub-headline, text, and graphics. The irregular block may contain images within texts. Vertical projection can be carried out to distinguish the images among text. In the evaluation process used fifty articles which have different format of paragraph arrangements and also include images. First, identified and got the count of regions manually. After that compared the result from identified regions and got the measurements. The region was identified with articles in the efficiency of 80.09%, headline extraction accuracy was 81.616%.

Keywords: archaeological records; image processing; headline extraction, histogram analysis.

[1] Corresponding author: Shanmugalingam Kasthuri. Tel.: +94-77-831-2321
E-mail address: s.shanshiya@gmail.com