59. NEWSPAPER LAYOUT SEGMENTATION: GETTING THE RELEVANT ADVERTISEMENTS

THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Special-Issue-27-December-2019

Newspaper Layout Segmentation: Getting the relevant Advertisements

Pooja Jain

Research Scholar, Dept. of Computer Science and Applications, Panjab University, Chandigarh

Kavita Taneja

Assistant Professor, Dept. of Computer Science and Applications, Panjab University, Chandigarh

Harmunish Taneja

Assistant Professor, Dept. of Computer Science and I.T., D.A.V. College, Sector- 10, Chandigarh.

Abstract:

India is one of the largest newspaper markets in the world with newspapers as a primary source for advertising jobs, tenders, admission notices and product sales and promotions etc. Search for relevant newspaper advertisements becomes very crucial for people waiting for such advertisements to be out in the newspapers. With the help of Internet technologies, online newspapers are becoming more and more popular, replacing printed version. Online newspapers are mostly available in .pdf formats for free download. Similar to printed version, searching for a specific advertisement in online newspapers also requires sequential manual search in multiple newspapers which is very time consuming and tedious. This paper presents an adaptive thresholding plus connected component based image processing technique to identify the images in the newspaper .pdf files. Non-advertisement images can be later filtered out and the relevant advertisement can be provided to the user after keyword matching.

Advertisements play a major role in our lives. Every day we get up in the morning with a cup of tea and newspaper in our hands, anxiously looking for particular advertisements including jobs, admission notices, tenders, sales, product launch, lost and found and much more. With the tremendous increase in the popularity and usage of internet equipped mobile devices during the last one decade, the focus has now shifted from printed newspapers to online newspapers (e-papers). So much so, that some newspapers are available only in electronic form and not as printed copies. Rather than waiting for the newspaper in the morning, people have started reading newspapers online (e-papers) as per their convenience (on their laptops or mobile phones) and according to their taste or need.

Along with the news articles, advertisements in the newspapers are of much interest.Many online newspapers give the advanced search options using which we can search particular news articles by giving the appropriate keywords but the same is not true with the advertisements in the newspapers. Also, no search engine including Google has a primary purpose of searching advertisements from online newspapers. As a result when we search for some advertisement in the newspapers through search portals, we may get hundreds of images but needless to say that only a few of them are relevant. Most of the times, news articles from newspapers containing the related keywords are served or old advertisements (which are of no relevance on the present day) are displayed. Their source is most of the times the Job portals or organization's websites but not the direct online newspapers.

To facilitate this advertisement search from online newspapers, newspaper layout segmentation needs to be performed to extract images from the online newspapers. Online newspapers are mostly available in .pdf formats for free download. Hence, the problem is identified as finding the images in these .pdf files as the first step.Theseimages can be further classified into various categories as per the needs of the user. This paper presents an adaptive thresholding plus connected component based image processing approach for newspaper layout segmentation.

The rest of the paper is arranged as follows. Section 2 reviews the literature for techniques used for newspaper layout segmentation.In section 3, an adaptive thresholding plus connected component based image processingalgorithm is proposed to separate out the images in the newspaper .pdf files. Section 4 explains the implementation of the proposed algorithm and section 5 presents the results and discussion followed by section 6 which concludes the paper along with the future scope of this research.

Literature Review

Before 2000, only rule-based approach for newspaper layout segmentation was used similar to Gatos et al. [1] (1999) where articles in the newspapers were identified by extracting image components like line, title blocks and image and drawing etc using rules. A dataset of 100 scanned pages of 'TO VIMA' newspaper S.A. was used here.

2000 onwards, bottom-up approach was used for newspaper layout segmentation. Liu et al. [2] (2001) proposed an algorithm for newspaper layout analysis using bottom-up approach in which connected components were detected first and then classified into basic components like line, text or graph components. Considering component attributes, basic components were merged by a heuristic rule.

Jatinder Kaur
Principal
Guru Gobind Singh College For Women
Sector 26, Chandigarh