

## INTISARI

Kompas Gramedia menerapkan proses pengarsipan koran-koran tua menjadi sebuah perpustakaan digital yang berisi artikel-artikel koran Kompas. Artikel-artikel tersebut diolah dan dikategorikan menjadi sebuah kumpulan data-data berbentuk teks artikel. Arsip-arsip ini pada mulanya direncanakan untuk kebutuhan internal instansi, seiring waktu arsip-arsip ini mulai digunakan untuk kepentingan umum.

Penelitian ini mencoba untuk membuat sebuah prototipe sistem segmentasi judul dan kolom teks artikel koran yang dapat membantu pengenalan karakter. Data masukan untuk sistem diperoleh dari hasil pemotongan citra koran secara manual dengan format \*.bmp. Jumlah data yang digunakan adalah 100 citra artikel koran dari periode 1980 dan 1990. Data *training* adalah 20 citra artikel dari tahun 1990 dan sisanya menjadi data *testing*. Kemudian sistem menyeleksi karakter-karakter menggunakan atribut-atribut dari konektivitas komponen, diperoleh judul dan kolom teks. Karakter-karakter judul dikelompokkan dan disegmentasi sesuai daerah barisnya menggunakan profil proyeksi. Citra masukan segmentasi kolom teks adalah hasil segmentasi judul yaitu citra artikel tanpa judul, kemudian akan disegmentasi dengan metode profil proyeksi untuk mensegmentasi kolom-kolom teks.

Maka dari hasil penelitian segmentasi judul dan kolom teks untuk artikel koran menggunakan metode profil proyeksi dan konektivitas komponen, diperoleh hasil untuk segmentasi judul dengan tingkat keberhasilan 91,25%. Segmentasi kolom teks secara vertikal meraih tingkat keberhasilan 100%. Pada beberapa artikel yang mengandung sub-judul menyebabkan sistem gagal mensegmentasi kolom teks secara horisontal.

## ABSTRACT

Kompas Gramedia archive the old newspapers into a digital library which contain newspaper article. The newspapers articles were processed and categorized into a collection of text data. These archives were initially necessary for internal agencies, then for a long time for public.

This research presented a prototype method of segmentation system that can help character recognition. The system *input* was obtained by cropping articles of newspaper and saved into \*.bmp formats compression. There were 100 images of newspaper article in 1980 and 1990 which was used for the research. The system was trained by 20 articles of 1990 and tested by 80 articles of 1980. Furthermore, characters connected component attributes were selected to distinguish between title character and text column character. Those characters were grouped(segmented) by their row region using projection profiles. The text column segmentation *input* was the result image of title segmentation, finally it would be segmented with projection profile.

The result, title segmentation achieved success rate 91.25%. The text column segmentation vertically achieved 100% success rates. Unfortunately, some articles contained sub-title caused the system failed in segmented the text column horizontally.