IMPLEMENTASI OPTICAL CHARACTER RECOGNITION (OCR) DENGAN PENDEKATAN METODE STRUKTUR MENGGUNAKAN EKSTRAKSI CIRI VEKTOR DAN REGION

Mirza Trilaksono[1], M. Ramdhani[2], Achmad Rizal[3]

[1]Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom

Abstrak

Optical Character Recognition (OCR) adalah sebuah sistem komputer yang digunakan secara otomatis mengenali serangkaian karakter yang berasal dari mesin ketik, mesin cetak ataupun tulisan tangan. Dengan kata lain OCR adalah proses pengalihan dokumen teks menjadi file komputer tanpa harus pengeditan ulang, setiap karakter baik huruf, kata, kalimat dapat dikenali secara tepat dan dibaca oleh perangkat lunak yang lain, tanpa harus pengetikan ulang dan editing.

Pada tugas akhir ini dikembangkan suatu aplikasi untuk mengidentifikasi karakter pada suatu file gambar (bmp) yang berisi karakter yang berasal dari pemindaian hardcopy atau dari sumber lainnya. Proses ekstraksi ciri menggunakan pendekatan vektor dan region. Pada proses tersebut akan ditentukan vektor penyusun garis karakter pada tiap area pengamatan (region), dimana tiap karakter dibagi menjadi 9 region yang sama besar dan simetris.

Untuk mengevaluasi performansi dari OCR dengan menggunakan metode tersebut, dilakukan pengujian terhadap beberapa sampel masukan baik yang berasal dari dokumen hardcopy maupn yang berasal dari sumber lainnya. Hasil analisis menunjukkan bahwa sistem OCR ini mempunyai tingkat akurasi sebesar 86,49% untuk font yang sudah dilatihkan, dan 63,35% untuk font yang belum dilatihkan.

Kata Kunci : Pengenal huruf otomatis , ekstraksi ciri , vektor , region

Abstract

Optical Character Recognition (OCR) is a computer system which is used automatically to recognize a part of character coming from typewriter, letterpress and or handwriting. In the other hand, OCR is a process of transferring the text document become the computer file without having to expurgation repeat, every characters such as letter, word, sentence can be recognized precisely and read by other software, without having to type repeating and editing.

In this final project will be developed an application to identify the character at one file picture (*.bmp) which contains character from hardcopy or other source. The extractions distinguish process using the approach of vector and region. At this process will be determined vector compiler mark with lines of character at every perception area, where each character divide into 9 regions that have same size and symmetries.

To evaluate the performance of OCR by using that method, will be conducted an examination to some input samples which is coming from document of hardcopy or other source. The result shows that this OCR system have recognition rate 86,49 % in trained font, and 63,35% in non trained font.

Keywords : Character recognition,feature extraction,vector,region

BAB I
PENDAHULUAN

1.1 Latar Belakang

Cukup banyak artikel-artikel *text* menarik yang disajikan dalam format gambar seperti *.jpg dan *.bmp. Sangat disayangkan sekali, andaikan artikel penting yang berlembar-lembar tersebut, ingin dijadikan sebagai sumber dari sebuah karya tulis harus diketik ulang seluruhnya. Hal tersebut cukup memakan waktu dan tenaga. Oleh karena itu diperlukan suatu teknik untuk 'mengkonversi' teks yang berformat gambar menjadi format *.txt agar dapat di-*copy* dan di-*edit*. Teknik ini menggunakan sistem *Optical Character Recognition* (OCR).

Optical Character Recognition (OCR) adalah sebuah sistem komputer yang dapat membaca huruf, baik yang berasal dari sebuah pencetak (printer atau mesin ketik) maupun yang berasal dari tulisan tangan. Dengan adanya sistem OCR maka user dapat lebih leluasa memasukkan data karena user tidak harus memakai papan ketik (*keyboard*) tetapi bisa menggunakan pena elektronik untuk menulis sebagaimana user menulis di kertas. Adanya pengenal huruf juga akan memudahkan penanganan pekerjaan yang memakai input tulisan seperti penyortiran surat di kantor pos, pemasukkan daftar nilai, dan pemasukkan data buku di perpustakaan.. Keberadaan sistem pengenal huruf yang cerdas akan sangat membantu usaha besar-besaran yang saat ini dilakukan banyak pihak, yakni usaha digitalisasi informasi dan pengetahuan, misalnya dalam pembuatan koleksi pustaka digital, koleksi sastra kuno digital, dll.

Pada tugas akhir ini penulis mencoba membangun suatu aplikasi OCR dengan menggunakan pendekatan vektor dan region pada ekstraksi cirinya. Diharapkan metode yang digunakan dapat dijadikan referensi sebagai salah satu metode untuk mengidentifikasi karakter yang handal yang memiliki tingkat akurasi lebih dari 75 persen.

1.2 Tujuan Penelitian

Tujuan yang ingin dicapai dalam tugas akhir ini adalah:

1. Membangun suatu aplikasi untuk mengidentifikasi karakter pada suatu file gambar yang berasal dari *hardcopy* dokumen atau dari sumber lainnya, dengan menggunakan pendekatan vektor dan region pada ekstraksi cirinya.
2. Menganalisis performansi aplikasi OCR dengan parameter tingkat keakuratan identifikasi.

1.3 Perumusan Masalah

Beberapa hal yang akan diteliti dalam Tugas Akhir ini yaitu:

1. Bagaimana proses ekstraksi ciri menggunakan pendekatan vektor dan region ?
2. Bagaimana proses pembelajaran terhadap input-input sampel karakter ?
3. Bagaimana proses pencocokan gambar input dengan sampel yang ada dalam database ?

1.4 Batasan Masalah

Batasan–batasan masalah yang digunakan dalam tugas akhir ini adalah:

1. Format file masukan dalam format BMP dan JPEG.
2. Implementasi sistem OCR tidak menggunakan blok *postprocessing* (*autospell* dan pengembalian format).
3. Text pada file gambar yang akan diinterpretasikan harus dalam posisi mendatar dan terpisah antar karakternya, serta bukan dalam format *italic*, *underline*, ataupun *strikethrough*.
4. Resolusi minimal file gambar adalah 200 ppi.
5. Implementasi menggunakan bahasa pemrograman Visual Basic 2005.

1.5 Metode Penyelesaian Masalah

Penelitian ini dilakukan dengan metodologi sebagai berikut:

1. Tahap studi literatur.

Studi literatur mengenai konsep-konsep pengenalan karakter dan pengolahan citra pada umumnya.

2. Tahap perancangan, realisasi perangkat.
   Perancangan sistem berdasarkan dari hasil studi literatur, pemodelan dari sistem tersebut diterjemahkan ke program simulasi dengan software Visual Basic 2005.
3. Tahap pengujian perangkat.
   Pada langkah ini akan diuji performansi dari aplikasi OCR yang telah dibuat.
4. Tahap analisis dan penarikan kesimpulan.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Pada bab I ini, dijelaskan mengenai latar belakang, tujuan, batasan masalah, dan metoda pelaksanaan penelitian serta sistematika pembahasan laporan.

BAB II DASAR TEORI

Bab ini merupakan tinjauan pustaka dari pengolahan citra, sistem OCR dan algoritma yang digunakan untuk implementasi sistem.

BAB III PERANCANGAN DAN IMPLEMENTASI PERANGKAT LUNAK

Perancangan dimulai dari deskripsi masalah dan persyaratan pengguna (*user requirements*). Pengembangan aplikasi, dan interpretasi algoritma dibahas di sini.

BAB IV PENGUJIAN DAN ANALISIS

Bab ini menguraikan pengujian dan analisis sistem. Evaluasi aplikasi OCR yang dihasilkan dibahas di sini. Beserta analisis performansi yang berhasil dicapai..

BAB V PENUTUP

Bab ini berisi simpulan dari implementasi yang dilakukan serta saran untuk pengembangan di masa mendatang.

BAB V
PENUTUP

5.1 Simpulan

1. Sistem OCR yang dibangun dengan pendekatan metode struktur menggunakan ekstraksi cirri vektor dan region memiliki tingkat akurasi sebesar 86,49% untuk font yang sudah dilatihkan, dan 63,35% untuk font yang belum dilatihkan.
2. Sistem OCR ini dapat bekerja dengan baik jika citra masukannya memiliki dimensi 200 ppi atau 300 ppi. Sistem OCR ini kurang dapat bekerja dengan baik untuk dimensi citra masukan 100 ppi karena dengan ukuran karakter yang terlalu kecil, proses segmentasi tidak berjalan dengan sempurna.
3. Makin besar dimensi citra masukan, maka makin lama waktu deteksi yang dilakukan sistem. Hal ini terjadi karena makin besar dimensi citra masukan, makin banyak pula jumlah piksel yang diproses.

5.2 Saran

1. Implementasi sistem OCR pada tugas akhir ini masih belum menggunakan blok postprocessing (*autospell* dan pengembalian format tulisan), sehingga untuk penelitian selanjutnya perlu mengintegrasikan blok ini agar tingkat akurasi sistem dapat meningkat.
2. Perlunya menggunakan algoritma segmentasi yang lebih baik agar sistem OCR dapat memisahkan karakter baik dalam format *italic*, *underline*, atau *strikethrough*.

DAFTAR PUSTAKA

Brown,Eric. 1992. *Character Recognition by Feature Point Extraction* (Online).
Tersedia : http://www.ccs.neu.edu/home/feneric/charrecpres.html (17 November 2007).

Haigh, Susan. 1996. *Optical Character Recognition (OCR) as a Digitization Technology*. Canada : National Library of Canada.

Hunn, Ketil. 2000. *Character Recognition Using Backpropagation In A Neural Network.* Pittsburgh : University of Pittsburgh.

Iyer, Singh, dkk. 2005. *Optical Character Recognition System for Noisy Images in Devanagari Script.* In UDL Workshop on Optical Character Recognition with Workflow and Document Summarization (OCR & DS-2005).

Kim, Thoma, dkk. 2000. *Automated Labeling in Document Images*. Bethesda : National Library of Medicine.

Nugraha, A.P. dan Mutiara, A.B. 2003. *Metode Ekstraksi Data Untuk Pengenalan Huruf Dan Angka Tulisan Tangan Dengan Menggunakan Jaringan Syaraf Buatan Propagasi Balik*. Depok : Universitas Gunadarma.

Srihari, Sagur, dkk. 2000. *Approximate Stroke Sequence String Matching Algorithm for Character Recognition and Analysis*. New York: Buffalo.

Thathoo Rahul, Suman Sekhar. 2000. *To Understand Is To Perceive Patterns.* Berlin : Isaiah Berlin.

Zhang, Waibel, dkk. 2000. *A PDA-based Sign Translator*.

Zhang Yungang, Zhang Changsui. 2000. *A New Algorithm for Character Segmentation of License Plate*. Beijing : Tsinghua University.