OPTICAL CHARACTER RECOGNITION (OCR) METODE STRUKTUR MENGGUNAKAN EKSTRAKSI KARAKTERISTIK TITIK DAN VEKTOR

Alex Kurniawan[1], Tjokorda Agung Budi Wirayuda[2], Retno Novi Dayawati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, 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 atau jpg) 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 91,88% untuk font yang sudah dilatihkan, dan 52,26% 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 or *.jpg) 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 91,88 % in trained font, and 52,26% in non trained font.

Keywords : Recognition , Extraction , vektor , region

1.PENDAHULUAN

1.1 Latar Belakang

Banyaknya 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) [1].

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. Optical Character Recognition (OCR) merupakan solusi untuk memudahkan usaha mendigitalisasikan informasi dan pengetahuan [1].

Salah satu metode dalam Optical Character Recognition (OCR) adalah metode struktur. Dalam metode yang berbasis struktur, setiap pola yang diproses dinyatakan sebagai gabungan beberapa struktur elementer. Proses pengenalan dilakukan dengan mencocokkan komposisi struktur elementer dengan struktur yang sudah disimpan memakai pendekatan karakteristik titik dan vektor. Sehingga memudahkan dan mempercepat dalam proses pengenalan dan pelatihan karakter.

Pendekatan karakteristik titik dan vektor merupakan metode untuk mengambil ciri dari karakter, sehingga dapat membedakan karakter yang satu dengan yang lain. Pendekatan ini dapat meningkatan akurasi pengenalan karakter dari metode struktur. Sehingga dapat menghasilkan Optical Character Recognition (OCR) dengan akurasi lebih dari 90 persen .

Pada tugas akhir ini penulis mencoba membangun suatu aplikasi OCR dengan menggunakan pendekatan ciri titik dan vektor pada ekstraksi cirinya. Diharapkan metode yang digunakan dapat dijadikan referensi sebagai salah satu metode untuk mengidentifikasi karakter yang handal, dan aplikasi OCR yang dihasilkan memiliki tingkat akurasi lebih dari 90 persen.

1.2 Perumusan Masalah

Masalah yang akan dibahas dalam Tugas Akhir ini yaitu:

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

Pembatasan Masalah

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

1. Format file masukan dalam format BMP dan JPG.
2. Text pada file gambar yang akan diinterpretasikan harus dalam posisi mendatar dan terpisah antar karakternya, serta bukan dalam format *italic*, *underline*, ataupun *strikethrough*.
3. Tidak menangani kemiringan text.
4. Menggunakan jenis font Times New Roman, MS Sans Serif dan Arial.
5. Resolusi minimal file gambar adalah 200 ppi.

1.3 Tujuan

Tugas Akhir ini bertujuan untuk :

1. Membangun suatu perangkat lunak untuk mengidentifikasi karakter pada suatu file gambar yang berasal dari hardcopy dokumen atau dari sumber lainnya, dengan menggunakan pendekatan berbasis struktur menggunakan karakteristik titik dan vektor pada ekstraksi cirinya.

2. Menganalisis performansi perangkat lunak OCR dengan parameter tingkat keakuratan identifikasi.

1.4 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang akan dilakukan dalam menyelesaikan tugas akhir ini adalah:

1. Studi literatur

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

2. Analisa dan Perancangan perangkat lunak.

   Menganalisis dan merancang perangkat lunak dengan pemrograman terstruktur dan membangun perangkat lunak dengan pendekatan berbasis struktur menggunakan karakteristik titik dan vektor.

3. Implementasi

   Mengimplementasikan sistem menggunakan bahasa pemrograman Visual Basic .NET dengan pendekatan berbasis struktur menggunakan karakteristik titik dan vektor.

4. Pengujian sistem

   Melakukan pengujian dengan mengambil sampel tulisan tangan 20 orang dan melakukan pengujian dengan perangkat lunak dengan pendekatan berbasis struktur menggunakan karakteristik titik dan vektor pada sistem.

5. Penyusunan laporan dan kesimpulan, menyusun laporan tugas akhir dan menarik kesimpulan akhir berdasarkan analisis yang dilakukan.

5. PENUTUP

5.1 Kesimpulan

1. Sistem OCR yang dibangun dengan pendekatan metode struktur menggunakan ekstraksi cirri vektor dan region memiliki tingkat akurasi sebesar 71.0254% untuk font yang sudah dilatihkan, dan 63.675% 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.
4. Makin besar ukuran font citra masukan, maka tinggi tingkat akurasi OCR. Hal ini terjadi karena makin besar ukuran font citra masukan, makin banyak pula jumlah piksel yang diproses.
5. Makin besar citra normalisasi terhadap objek citra dalam proses ekstraksi mampu mempengaruhi tingkat keseragaman dalam hal ketebalan dan ukuran menjadi lebih baik
6. Makin banyaknya pembagian matrik citra dalam proses ekstraksi mampu mempengaruhi tingkat akurasi menjadi lebih baik
7. Penyusunan tahap-tahap preprocessing mampu mempengaruhi tingkat keseragaman dalam hal ketebalan dan ukuran. Sehingga mempengaruhi tingkat perolehan akurasi.

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

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

[2] Das, Koushik. 2000. *Design and Implementation of an Efficient Thinning Algorithm* [on-line]. Indian Institute of Technology Kanpur; tersedia dari http://citeseer.nj.nec.com/cache/papers/cs/25859/http:zSzzSzwww.cse.iitk.ac.inzSzresearchzSzbtp2000zSzkdas.pdf/das00design.pdf; Internet; diakses pada 14 November 2007

[3] Fisher, Bob, Simon Perkins, Ashley Walker, and Erik Wolfart. 1994. *Skeletonization/Medial Axis Transform* [on-line]. University of Edinburgh; tersedia dari http://www.cee.hw.ac.uk/hipr/html/skeleton.html; Internet; diakses pada 15 November 2007

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

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

[6] 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).

[7] ITU Standaritation. 1999. Grayscale: London.

[8] J.J. Hull, "Database for handwritten word recognition research" IEEE PAMI 16 1994, 550-554

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

[10] Nalwan, Agustinus, *Pengolahan Gambar Secara Digital*, Elex Media Komputindo, Jakarta (1997).

[11] Petroutsos, Evangelos. 1996. *Mastering Visual Basic 6.* SYBEX Inc, 1151 Marina Village Parkway, Alameda, CA 94501 : San Francisco London.

[12] Pressman, Roger S.2001. *Software Engineering*. McGraw-Hill Higher Education : New York.

[13] Rinaldi, Munir. 2004. *Pengolahan Citra Digital Dengan Pendekatan Algoritma.* Informatika : Bandung.

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

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

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

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