ISSN: 2089-1121

OPTICAL CHARACTER RECOGNITION DAN INFORMATION EXTRACTION PADA DOKUMEN KAMUS BILINGUAL

F.X. Ferdinandus*), Arya Tandy Hermawan*), Melisa Tedjokusumo), dan Lukman Zaman*)**
*) Program Pascasarjana Teknologi Informasi
Sekolah Tinggi Teknik Surabaya
**) Jurusan Teknik Informatika
Sekolah Tinggi Teknik Surabaya
ferdi@stts.edu , arya@stts.edu , tmmt_89@yahoo.com , luqmanz@gmail.com

ABSTRAK

Perkembangan teknologi memungkinkan dilakukannya otomatisasi terhadap berbagai kegiatan, salah satunya ialah dalam memperoleh informasi dari dokumen hasil cetak. Untuk itu, dibuat suatu sistem yang mampu mengenali dokumen hasil cetak yang dalam hal ini merupakan kamus bilingual dan mendapatkan informasi dari hasil pengenalan tersebut. Sistem ini terdiri atas dua bagian utama yaitu Optical Character Recognition (OCR) dan Information Extraction (IE).

Proses OCR terdiri atas proses preprocessing dan recognition menggunakan metode Combined Template Matching. Metode combined template matching ini merupakan metode dasar OCR yaitu Template Matching yang dikombinasikan dengan penggunaan Induksi Decision Tree. Kemudian hasil pengenalan karakter tersebut akan diekstrak menggunakan IE untuk memisahkan dokumen kamus ke dalam field-field database berdasarkan strukturnya.

Hasil dari proses OCR dipengaruhi oleh banyak faktor, seperti usia dokumen, resolusi scanner, sudut kemiringan dan sebagainya. Oleh karena itu proses OCR memerlukan tahap preprocessing untuk mempersiapkan citra sebelum dapat dikenali. Sedangkan pada IE, diperlukan informasi struktur record karena setiap kamus bilingual yang berbeda dapat memiliki perbedaan struktur penulisannya. Dalam aplikasi ini akurasi dari proses Optical Character Recognition mencapai kisaran 80 sampai 90%.

Kata kunci: optical character recognition, template matching, induksi decision tree, information extraction.

ABSTRACT

The development of today's technology allows for the automation of various activities, one of them is to gaining information from printed document. For that, a system that is capable of recognizing printed document, in this case bilingual dictionary documents and gaining structured information from them is established. This system consists of two parts, the Optical Character Recognition (OCR) and Information Extraction (IE).

The OCR process consists of preprocessing and recognition process that uses Combined Template Matching method. This is a basic OCR method, made by combining Template Matching with the use of Decision Tree Induction. The result of character

ISSN: 2089-1121

recognition is extracted using IE to separate each field of the dictionary document to database fields based on the structure.

The result of OCR process are influenced by many factors, such as document age, scanner resolution, angle declivity, et cetera. Therefore, the OCR process needs preprocessing steps to prepare an image before it can be recognized. Whereas the IE process needs structured information records because documents of different bilingual dictionaries may have different writing structures. The accuracy of Optical Character Recognition process in this application may vary between 80% to 90%.

Keywords: optical character recognition, template matching, decision tree induction, information extraction.

1. PENDAHULUAN

Optical Character Recognition (OCR) merupakan salah satu sub bidang dari *Computer Vision* yang bertujuan untuk melakukan pengenalan terhadap karakter-karakter hasil cetak. OCR banyak digunakan untuk mengubah dokumen atau buku menjadi file elektronik, sehingga memungkinkan untuk mengedit teks, mencari kata atau frase, menyimpan, menampilkan atau mencetak salinan dari teks hasil scanning, dan sebagainya. Sub bidang OCR ini dapat dimanfaatkan untuk melakukan komputerisasi terhadap pekerjaan-pekerjaan yang masih manual.

Menyalin suatu dokumen atau buku akan memerlukan waktu yang lama dan melelahkan. Dengan bantuan komputer dan menggunakan teknologi optical character recognition, pekerjaan tersebut dapat dilakukan dengan lebih cepat dan juga lebih mudah. Selain itu, dengan memanfaatkan ketelitian yang dimiliki oleh komputer, diharapkan dapat meminimalisir *human-error* yang sering terjadi pada pekerjaan yang dilakukan secara manual.

Selain itu, seiring dengan berkembangnya teknologi informasi, kebutuhan akan informasi turut meningkat. Oleh karena itu muncul pula teknologi yang bertujuan mendapatkan informasi dari suatu dokumen, yaitu teknologi *Information Extraction*.

Information extraction merupakan teknologi yang disediakan oleh salah satu subbidang kecerdasan buatan, yaitu *Natural Language Processing*. Natural Language Processing merupakan subbidang *computer science* yang terkait dengan hubungan antara komputer dan bahasa alami manusia. Information extraction ialah teknologi yang berupaya mendapatkan informasi terstruktur dari suatu dokumen yang tidak/kurang terstruktur.

2. TEMPLATE MATCHING

Optical Character Recognition merupakan proses untuk mengubah atau mengkonversi suatu dokumen hasil cetak seperti buku, majalah, literatur atau surat kabar secara otomatis sehingga karakter-karakter yang terdapat pada dokumen tersebut dapat dikenali oleh komputer. Secara umum, suatu sistem kerja optical character recognition dapat terdiri atas berbagai proses, antara lain *data capture*, preprocessing, segmentasi, recognition dan postprocessing.

Salah satu metode OCR adalah Template Matching. Template matching merupakan suatu metode optical character recognition dilakukan dengan cara mencari template yang paling cocok dengan suatu image atau mencocokan suatu glyph dengan template master. Pencocokan antara glyph dengan template ini dilakukan pixel per pixel untuk mengetahui pixel mana saja yang sama dan pixel mana yang berbeda antara glyph

ISSN: 2089-1121

dengan templatenya. persamaan atau perbedaan pixel inilah yang kemudian dijadikan sebagai acuan oleh metode-metode template matching dalam mencari nilai kecocokan antar glyph dengan templatenya. Dalam aplikasi ini digunakan tiga metode template matching yaitu Square Different Matching (SQDIFF), Correlation Matching (CCORR), dan Correlation Coefficient Matching (CCOEFF).

Pada proses OCR dengan metode combined template matching, digunakan decision tree untuk membantu menentukan kecocokan. Decision tree merupakan suatu alat untuk membantu membuat keputusan yang digambarkan/dimodelkan dalam bentuk seperti pohon (tree). Decision tree digunakan untuk membantu mengidentifikasi strategi yang paling cocok untuk mencapai suatu tujuan. Induksi decision tree termasuk dalam sub bidang *Machine Learning* yaitu *Data Mining* dan digunakan untuk melakukan klasifikasi data.

Untuk mendapatkan informasi terstruktur dari dokumen yang telah dikenali, dilakukan proses Information Extraction. Information Extraction merupakan sebuah jenis pencarian informasi yang bertujuan mendapatkan informasi terstruktur secara otomatis dari dokumen yang tidak/kurang terstruktur.

3. OPTICAL CHARACTER RECOGNITION UNTUK KAMUS BILINGUAL

Optical Character Recognition merupakan suatu proses yang bertujuan mengkonversi suatu dokumen hasil cetak seperti buku, majalah, literatur atau surat kabar sehingga karakter-karakter yang terdapat pada dokumen tersebut dapat dikenali oleh komputer. Suatu sistem OCR terdiri atas beberapa proses, antara lain preprocessing, segmentasi, recognition. Dalam proses image analysis, preprosessing dan segmentasi dianggap sebagai dua proses yang terpisah. Namun dalam aplikasi ini, segmentasi dianggap termasuk dalam preprocessing, karena segmentasi merupakan salah satu tahapan awal dalam mempersiapkan citra.

Dalam melakukan pengenalan karakter, digunakan suatu library yang berfokus pada computer vision yaitu OpenCV. OpenCV (*Open Computer Vision*) merupakan sebuah library *open source* yang berfokus pada bidang computer vision. Library ini dibuat pertama kali oleh tim Intel Performance Primitives dan dirilis pada tahun 1999, namun karena merupakan library open scource yang banyak digunakan baik untuk komersial maupun dalam bidang penelitian.

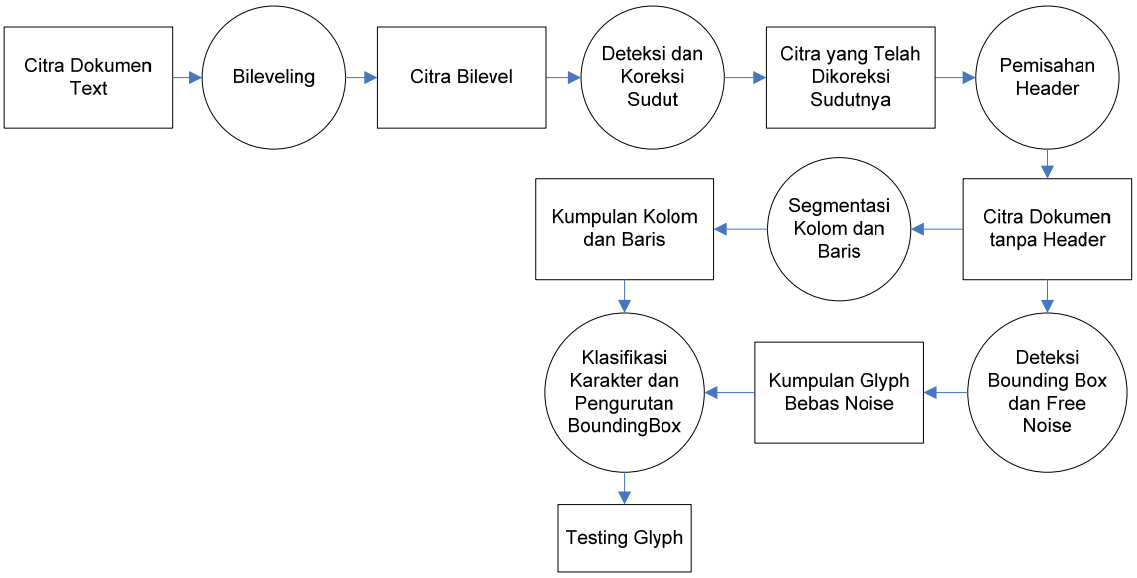
Dalam proses pengenalan karakter (OCR), preprocessing dilakukan untuk mempersiapkan suatu citra dokumen agar dapat digunakan sebagai input dalam proses selanjutnya, yaitu proses recognition atau pengenalan karakter. Sebelum melakukan preprocessing, dokumen hasil cetak yang akan diproses perlu diubah terlebih dahulu menjadi citra digital dengan cara melakukan scanning.

Dokumen yang diproses dalam aplikasi ini merupakan dokumen yang hanya berupa text yang disebut dengan *textual document image*. Preprocessing terdisi atas bileveling, koreksi sudut, pemisahan header, deteksi kolom dan baris, deteksi bounding box dan pembersihan noise, serta klasifikasi karakter dan pengurutan bounding box yang digambarkan pada arsitektur sistem OCR. Sistem OCR terdiri atas preprocessing dan recognition menggunakan combined template matching. Arsitektur sistem preprocessing dapat dilihat pada gambar 1.

Setelah tahap-tahap preprocessing selesai dilakukan, berikutnya dilakukan proses recognition atau pengenalan karakter. Input dari proses recognition adalah kumpulan glyph yang telah dihasilkan dari proses preprocessing yang selanjutnya akan disebut testing glyph. Sedangkan outputnya ialah file berisi karakter-karakter yang telah

ISSN: 2089-1121

dikenali yang ditulis dalam format html. Terdapat beberapa metode yang dapat digunakan dalam proses recognition, Dua metode yang paling umum digunakan dalam proses recognition ialah *feature extraction* dan *template matching*.



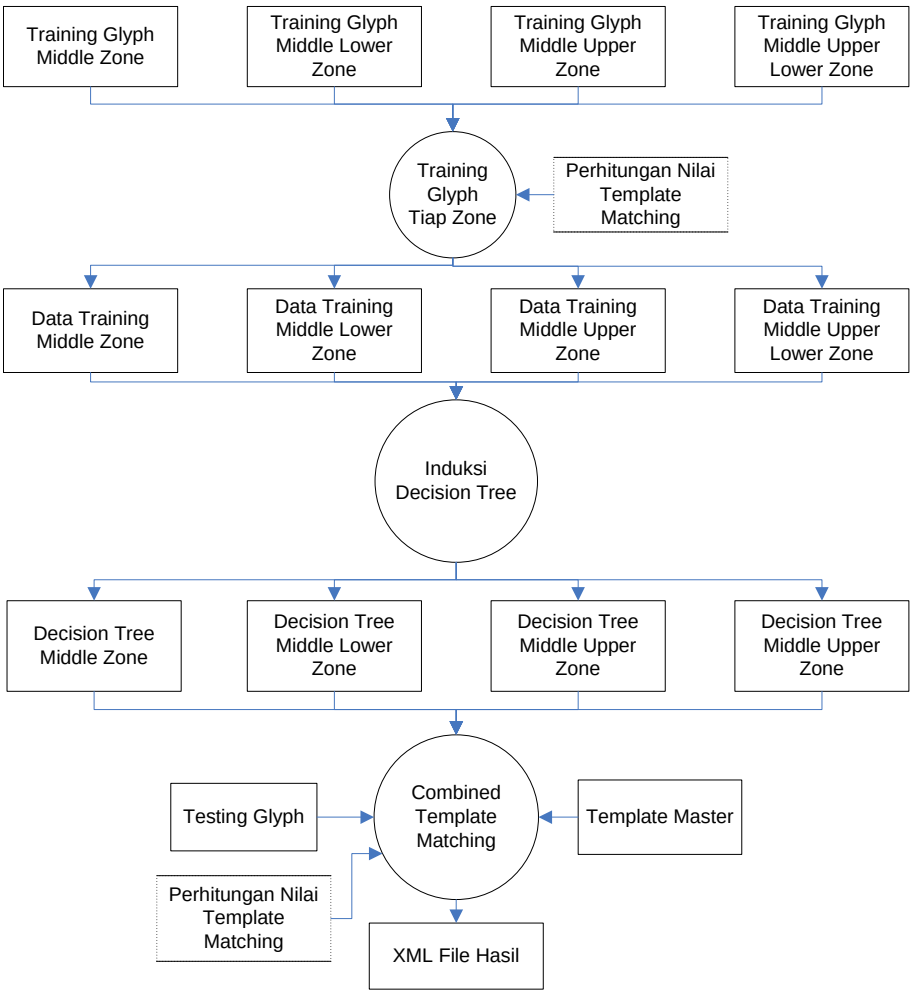
Gambar 1. Arsitektur Sistem Preprocessing

Pada aplikasi ini, metode pengenalan karakter yang digunakan ialah metode *Combined Template Matching*. Combined template matching merupakan pengembangan dari metode template matching yang mengkombinasikan metode template matching dengan penggunaan induksi decision tree. Penggunaan decision tree ini berfungsi menentukan nilai threshold yang akan digunakan dalam memutuskan cocok tidaknya suatu karakter dengan suatu template. Porses combined template matching terdiri atas tiga bagian yaitu training glyphs, induksi decision tree, dan testing glyph seperti yang digambarkan pada gambar 2.

Proses template matching merupakan suatu proses yang membutuhkan waktu cukup lama karena harus membandingkan satu per satu pixel pada glyph yang akan dikenali dengan pixel pada template-template yang digunakan. Oleh karena itu digunakan pembagian zone untuk memisahkan template-template yang akan digunakan serta decision tree untuk menentukan kecocokan glyph dengan template sehingga glyph tidak perlu dibandingkan dengan seluruh template yang ada.

Training glyph dilakukan untuk masing-masing zone yang diklasifikasikan berdasarkan *Board Clasification*. Hasil training glyph tersebut digunakan sebagai input dalam pembentukan decision tree. Decision tree yang telah terbentuk disimpan ke dalam suatu file XML. Decision tree tersebut kemudian digunakan dalam memprediksi glyph-glyph dari dokumen yang sedang diproses. Prediksi dilakukan dengan cara membandingkan hasil template matching antara glyph dengan template master, dengan nilai threshold yang didapat dari pelatihan glyph. Hasil yang didapatkan dari prediksi decision tree berupa cocok atau tidaknya glyph dengan template.

ISSN: 2089-1121



Gambar 2. Arsitektur Sistem Combined Template Matching

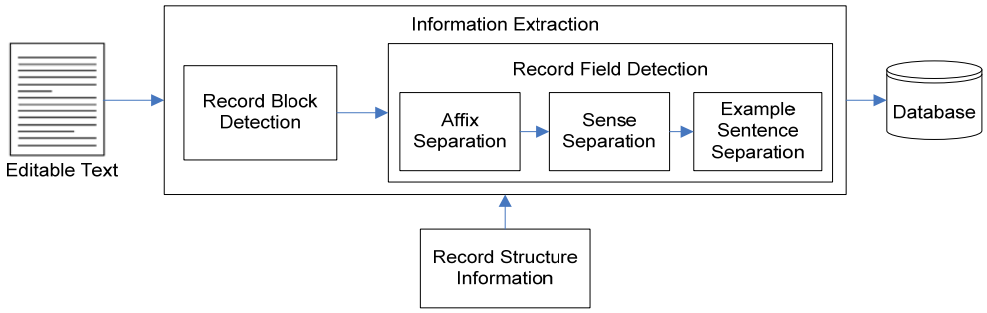
Apabila glyph diprediksi sudah cocok dengan template yang dibandingkan, glyph tersebut tidak akan dibandingkan dengan template lainnya melainkan langsung memproses glyph berikutnya. Glyph-glyph yang telah berhasil dikenali kemudian disimpan dalam suatu file yang ditulis dalam format penulisan HTML/XML.

4. INFORMATION EXTRACTION

Information Extraction (IE) adalah salah satu teknologi yang disediakan oleh Natural Language Processing untuk mengekstrak informasi yang terstruktur dari suatu dokumen yang tidak/kurang terstruktur. Natural Language Processing merupakan subbidang computer science yang terkait dengan hubungan antara komputer dan bahasa alami manusia. Information extraction dilakukan untuk menemukan hubungan antar entitas. Aplikasi ini akan memanfaatkan information extraction untuk membantu dalam melakukan penalaran logis untuk mencari kesimpulan misalnya dalam proses membedakan fungsi/kegunaan karakter-karakter tertentu.

Proses information extraction dalam aplikasi ini terdiri atas dua proses utama yaitu proses pemisahan blok dan proses pemisahan field. Dalam kamus bilingual sebuah blok terdiri atau sebuah kata dasar dimana kata dasar tersebut dapat memiliki banyak arti, kategori, cara baca, kata bentukan, serta contoh kalimat.

ISSN: 2089-1121



Gambar 3. Arsitektur Sistem Information Extraction

Proses information extraction diawali dengan meminta inputan struktur dokumen dari user. Kemudian dilakukan pemisahan blok dalam dokumen. Dari setiap blok yang berhasil dipisahkan, dipisahkan lagi berdasarkan field-fieldnya. Proses pemisahan field itu sendiri terdiri atas beberapa tahap, yang pertama yaitu memisahkan kata bentukan. Kata dasar sendiri akan digolongkan sebagai salah satu kata bentukan. Kemudian setiap kata bentukan dapat memiliki satu cara baca, satu kategori, banyak arti kata, dan banyak contoh kalimat.

5. UJI COBA

Uji coba ini dilakukan pada tiga buah buku kamus bilingual dimana salah satunya terdiri atas dua bagian kamus. Implementasi uji coba ini telah dilakukan pada 160 halaman kamus bilingual. Setiap halaman discan dengan resolusi 300 dpi dan skala 100% yang disimpan ke dalam file bitmap. Yang akan dijelaskan pada uji coba ini adalah hasil testing glyphs.

Dalam proses testing glyphs, glyph-glyph dalam dokumen dicocokan dengan template master untuk diketahui nilai kecocokannya. Nilai tersebut kemudian dimasukkan sebagai atribut dalam prediksi decision tree. Decision tree kemudian akan menilai apakah glyph tersebut sudah cocok dengan template master yang sedang diujikan. Apabila dinilai cocok, maka proses testing akan dilanjutkan ke glyph berikutnya. Dari proses pengenalan karakter ini, didapatkan rata-rata akurasi sekitar 80,79 %. Pada hasil combined template matching tersebut, terdapat cukup banyak karakter yang gagal dikenali. Hal ini dikarenakan threshold yang dihasilkan decision tree tidak mencakup karakter-karakter yang gagal dipisahkan, sedangkan pada saat proses testing glyph terdapat banyak kaarakter yang tergabung dan dipisahkan menggunakan pendekatan histogram projection.. Oleh karena itu dilakukan pengecekan ulang terhadap karakter-karakter yang gagal dikenali menggunakan salah satu metode template matching yang dianggap memiliki akurasi paling baik, yaitu CCOEFF.

Dengan melakukan teknik ini, banyak karakter yang sebelumnya gagal dikenali berhasil dikenali. Dengan demikian, persentase keberhasilan program dapat meningkat hingga mencapai kisaran 88,22%. Namun dari hasil tersebut, masih sering terdapat beberapa kesalahan pengenalan yang berupa kesalahan format seperti huruf regular dikenali sebagai huruf italic atau huruf regular dikenali sebagai huruf bold dan sebaliknya. Untuk itu dilakukan sebuah proses tambahan yaitu postprocessing dimana proses ini bertujuan memperbaiki format karakter-karakter yang salah dikenali formatnya berdasarkan karakter-karakter lain di sekitarnya. Pada tabel 1 ditunjukkan perbandingan hasil awal uji coba pengenalan karakter, uji coba setelah dilakukan perbaikan tingkat akurasi, serta uji coba setelah ditambahkan postprocessing.

ISSN: 2089-1121

Tabel 1 Hasil Uji Coba Pengenalan Karakter (a) Uji Coba Awal, (b) Uji Coba dengan Perbaikan Tingkat Akurasi, (c) Uji Coba dengan Postprocessing

(a)

Jumlah Karakter	Jumlah Karakter		Persentase Keberhasilan
	Salah Dikenali	Gagal Dikenali	
97	9	8	84,5%
144	12	15	81,25%
142	15	15	78,9%
41	5	3	80,5%
33	4	3	78,8%

(b)

Jumlah Karakter	Jumlah Karakter		Persentase Keberhasilan
	Salah dikenali	Gagal dikenali	
97	4	8	87,6%
144	9	15	83,3%
142	12	15	80,9%
31	3	3	80,6%
33	4	3	78,8%

(c)

Jumlah Karakter	Jumlah Karakter		Persentase Keberhasilan
	Salah dikenali	Gagal dikenali	
97	9	0	90,7%
144	8	0	94,4%
142	20	2	84,5%
41	8	0	80,5%
33	3	0	90,9%

Dari hasil pada tabel tersebut, didapatkan rata-rata akurasi program meningkat menjadi 91,32%. Sehingga dapat disimpulkan bahwa koreksi format yang ditambahkan pada postprocessing dapat meningkatkan akurasi program hingga 3,1%.

6. PENUTUP

Berdasarkan hasil uji coba yang telah dilakukan, dapat diambil kesimpulan sebagai berikut. Pengenalan karakter pada dokumen kamus bilingual membutuhkan perhatian khusus. Hal ini disebabkan karena ukuran karakter yang kecil, jarak antar baris yang saling berdekatan, serta banyaknya jenis karakter yang digunakan dimana terdapat banyak karakter yang memiliki banyak kemiripan. Pada proses information extraction, diperlukan informasi struktur record dari user karena setiap dokumen kamus bilingual dapat memiliki struktur yang berbeda-beda.

Sedangkan dari hasil uji coba yang dilakukan, dapat diberikan beberapa saran untuk pengembangan aplikasi ini. Penambahan template baru untuk karakter-karakter yang seringkali gagal/salah dikenali dapat memperbaiki akurasi dari aplikasi ini.

7. DAFTAR PUSTAKA

Agam, Gady, 2006, *Introduction to programming with OpenCV*, Illinois Institute of Technology.

Aryani, Iin. *Pengertian Optical Character Recognition*. http://ilmucerdas.wordpress.com/profil/pengertian-optical-character-recognition.

G. Bradski dan A. Kaehler. *Learning OpenCV*. O'Reilly. 2008.

Gunawan. *Pengantar ke Artificial Intelligence.* (Materi Kuliah Kecerdasan Buatan semester genap 2009/2010, Sekolah Tinggi Teknik Surabaya).

Lukmanto, Cristine Siandawati. *Optical Character Recognition Berbasis Combined Template Matching.* Sekolah Tinggi Teknik Surabaya, Surabaya. 2000.