**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

1. Identitas Pengusul

Nama : Erliyah Nurul Jannah

NRP : 5108100710

Dosen wali : Bilqis Amaliah, S.Kom, M.Kom

1. Judul Tugas Akhir

***“Rekonstruksi Dokumen Teks Arab yang Terdistorsi Geometris pada saat Akuisisi Citra”***

1. Latar Belakang

Akhir-akhir ini, digitalisasi dokumen kerap kali dilakukan oleh perpustakaan-perpustakaan di berbagai belahan dunia. Digitalisasi adalah proses penyimpanan dokumen dalam bentuk digital atau *soft copy*, bukan dalam bentuk fisik atau lembaran-lembaran kertas. Jika para penggemar buku ingin membaca buku digital tersebut, maka mereka bisa melakukannya dengan bantuan komputer.

Digitalisasi dibutuhkan untuk menyimpan dokumen-dokumen lama yang hampir rusak, yang tidak memungkinkan lagi untuk disimpan dalam bentuk lembaran-lembaran kertas. Karena lembaran-lembaran tersebut lambat laun akan melapuk. Tetapi tidak menutup kemungkinan bagi dokumen yang baru dicetak untuk dibuat versi digitalnya. Karena dengan versi digital ini dapat mengurangi biaya percetakan.

Islam mempunyai banyak dokumen penting yang sudah lama keberadaannya dan perlu dilestarikan agar tidak rusak. Dokumen-dokumen tersebut sebagian besar ditulis dalam bahasa arab. Sekarang ini, digitalisasi dokumen arab sangat diperlukan mengingat banyak dokumen-dokumen arab yang tidak mendapat perwatan yang layak.

Proses penyimpanan dokumen dalam bentuk digital dapat dilakukan dengan cara memindai dokumen tersebut menggunakan alat pindai atau *scanner*. Untuk buku dengan jumlah halaman yang sedikit, *scanner* tidak akan mengalami kendala dalam proses pemindaian. Namun jika buku tersebut memiliki halaman yang tebal maka pemindaian akan sulit dilakukan. Karena buku tersebut akan dipaksa dalam posisi terbuka sampai semua tulisan yang ada dalam tiap halaman dapat dikenali oleh *scanner*. Hal ini dapat merusak jilid buku.

Selain dengan menggunakan *scanner*, digitalisasi juga dapat dilakukan dengan menggunakna kamera digital. Namun, kendala yang terjadi pada *scanner* juga terjadi pada kamera digital. Yakni jika dokumen yang akan dibuat versi digitalnya tersebut memiliki halaman yang tebal. Selain dapat merusak jilid buku, dokumen yang tebal dapat membuat gambar hasil pemindaian terlihat melengkung permukaan halamannya.

Digitalisasi dokumen dengan *scanner* atau kamera digital pada umumnya bertujuan untuk menyimpan informasi tulisan atau teks yang dimuat dalam dokumen tersebut. Untuk mendapatkan teks dari dokumen hasil pindaian dapat dilakukan dengan bantuan *Optical Character Recognition* (OCR). OCR adalah suatu perangkat lunak yang dapat mengenali tulisan dalam suatu *file* gambar.

Namun, tidak semua gambar dokumen yang diambil dengan kamera digital atau *scanner* dapat dikenali oleh OCR. Dalam artian OCR dapat menampilkan isi tulisan yang ada dalam gambar dokumen. Baik menggunakan *scanner* maupun kamera digital, gambar dokumen yang dihasilkan dari buku-buku tebal biasanya memiliki kualitas yang kurang baik. Biasanya gambar permukaan halaman yang dihasilkan terlihat melengkung atau mengalami distorsi geometris. Distorsi ini membuat OCR tidak dapat mnegenali teks yang ada dalam citra dokumen.

Beberapa metode talah diajukan untuk menangani masalah tersebut. Yaitu agar gambar dokumen hasil kamera digital atau *scanner* yang permukaannya melengkung dapat dikenali oleh OCR. Metode-metode ini dapat dikelompokkan menjadi dua kategori. Yang pertama adalah metode yang berbasis *hardware*, dan yang kedua berbasis *software*.

Metode yang berbasis *hardware* misalnya adalah memotret dokumen dengan menggunakan kamera stereo, peralatan fotometrik, atau *laser scanner*. Metode yang berbasis *hardware* ini membutuhkan biaya yang lebih banyak, Karena harga dari peralatan memotret tersebut tidaklah murah. Sehingga metode ini tidak banyak yang mengembangkan.

Para peneliti lebih tertantang untuk mengembangkan metode-metode yang berbasis *software*. Metode-metode tersebut di antaranya adalah *Continuous Skeletal Image Representation* (SKEL)[1], *Segmentation Based Recovery* (SEG)[2], dan *Coordinate Translation Model* (CTM)[3].

SKEL sukses dalam mengubah permukaan yang melengkung menjadi datar, namun tidak berhasil dalam memperbaiki *perspective distortion*. Berbeda dengan SKEL, CTM tidak berhasil dalam memperbaiki permukaan yang melengkung maupun *perspective distortion*. Namun CTM mampu merestorasi rumus dan komponen lain milik halaman tetangganya yang ikut tertangkap oleh kamera. SEG lebih baik dari kedua metode tersebut, karena metode ini mampu memperbaiki permukaan yang melengkung dan permasalahan komponen lain milik halaman tetangga yang ikut terambil[4].

Semua metode tersebut dirancang untuk menangani permasalahan pada dokumen latin. Jika metode tersebut diterapkan pada dokumen arab, maka akan menghadapi masalah yang lebih kompleks. Karena tulisan arab berbeda dengan tulisan latin. Huruf arab mempunyai tinggi dan kemiringan yang tidak teratur. Selain itu, susunan tulisan latin lebih terorganisir dibanding tulisan arab. *Noise* pada dokumen arab juga lebih banyak dari pada dokumen latin [4].

Oleh sebab itu, dalam tugas akhir ini, metode baru yang merupakan perbaikan metode SEG[2] akan digunakan untuk merekonstruksi citra dokumen arab. Agar dokumen arab yang mempunyai permukaan melengkung dapat diperbaiki menjadi memiliki permukaan yang datar. Sehingga dokumen arab tersebut dapat diekstrak tulisannya menggunakan *Arabic* OCR.

1. Rumusan Masalah

Permasalahan dalam tugas akhir ini dapat dirumuskan sebagai berikut :

1. Bagaimana mereduksi *noise* yang terdapat dalam citra dokumen arab?
2. Bagaimana mendeteksi kemiringan tulisan dalam citra dokumen arab ?
3. Bagaimana memperbaiki kemiringan tulisan dalam citra dokumen arab?
4. Batasan Masalah

Tugas akhir ini mempunyai batasan masalah sebagai berikut :

1. Distorsi geometris yang dimaksud dalam tugas akhir ini berupa permukaan halaman yang melengkung
2. Tujuan

Merekonstruksi citra dokumen teks arab yang terdistorsi secara geometris pada saat akuisisi citra.

1. Manfaat

Manfaat dari tugas akhir ini adalah :

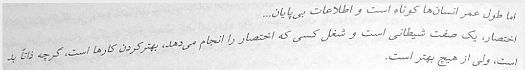
1. Dapat mereduksi *noise* yang terdapat dalam citra dokumen arab
2. Dapat mendeteksi kemiringan tulisan dalam citra dokumen arab
3. Dapat memperbaiki kemiringan tulisan dalam citra dokumen arab
4. Ringkasan Tugas Akhir

Citra dokumen tebal yang diambil menggunakan kamera digital atau *scanner* kadang kala mengalami ditorsi geometris. Distorsi geometris yang dimaksud adalah kerusakan pada permukaan halaman dokuma, misalnya permukaan yang terlihat melengkung. Contoh citra dokumen yang terdistorsi geometris dapat dilihat pada Gambar 1. Citra dokumen yang seperti ini tidak dapat dikenali oleh OCR (Optical Character Recognition).

Dalam mengubah permukaan melengkung pada dokumen arab menjadi permukaan yang datar (*flat*) terdapat langkah-langkah sebagai berikut :

1. Binarisasi citra

Binarisasi citra adalah mengubah citra berwarna atau hitam putih menjadi citra biner, yaitu citra yang mempunyai nilai piksel 0 dan 1. Binarisasi citra dapat dilakukan dengan menerapkan metode *thresholding* tertentu. Dalam tugas akhir ini, metode *thresholding* yang digunakan adalah metode Otsu.

**Gambar 1. Citra dokumen arab dengan permukaan melengkung**

1. Reduksi *Noise*

Setelah citra input diubah menjadi citra biner, selanjutnya adalah mengurangi atau mereduksi *noise* yang ada dalam citra. Reduksi *noise* ini bertujuan untuk mengurangi piksel-piksel yang bukan termasuk teks arab. Karena piksel-piksel ini tidak dibutuhkan dalam proses selanjutnya.

Reduksi noise dapat dilakukan dalam beberapa tahap, yaitu :

1. Ekstraksi *bounding box* dari tiap-tiap kata
2. Membuat histogram dari tinggi *bounding box* yang telah didapat sebelumnya
3. Mencari *bounding box* yang mempunyai frekuensi tertinggi, labeli dengan *M*
4. Melakukan *Low Pass Filtering* dengan cara mengeliminasi *word* atau kata yang memiliki tinggi lebih besar dari *a.M*, kurang dari *M.b*, atau kata yang memiliki lebar kurang dari *M/c* dengan *a,b,c* adalah koefisien yang didapat melalui *empirical practice*.
5. Deteksi *Word Box* dan *Text Line*

Deteksi *word box* dapat dilakukan dengan langkah-langkah sebagai berikut :

1. Melakukan *Top-Down Scanning* pada citra
2. Baris pertama atau baris yang paling atas pada teks dilebeli dengan *L*, sedangkan kata pertama pada baris *L* tersebut dilabeli dengan *K*
3. Untuk tiap-tiap baris pada teks, lakukan *right left scanning*
4. Kata di sebelah kiri *K* yang mempunyai jarak paling dekat dengan *K* akan dilabeli dengan *Kf*. *K* dan *Kf*ini dideteksi sebagai satu *word* atau kata karena mempunyai jarak yang dekat.

Sedangkan untuk deteksi *Text Line* dapat dilakukan dengan cara mendeteksi *upper* dan *lower baseline* dari tiap kata dengan menggunakan rumus : *y = aijx + bij* dan *y = a’ijx + b’ij* dengan *aij, bij, a’ij, b’ij* adalah koefisien yang didapat berdasarkan karakteristik dari kata.

1. Deteksi *Word Slope*

*Slope* atau kemiringan dari tiap-tiap kata dapat diturunkan dari *slope baseline* yang bersesuaian dengan rumus  *= arctan (aij)* untuk *upper baseline* dan  *= arctan (a’ij)* untuk *lower baseline*.

Dengan menggunakan estimasi kemiringan dari tiap kata, sudut antara kata dengan garis horizontal dapat diketahui dengan rumus :

untuk kata pertama pada tiap baris

untuk kata selain kata pertama.

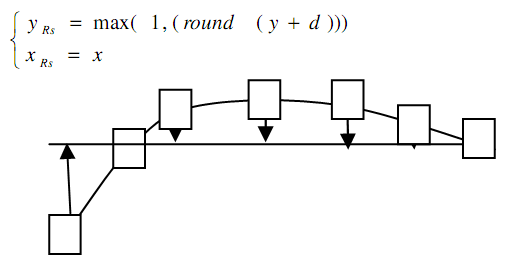
1. Koreksi *Word Skew*

Koreksi word skew dapat dilakukan dengan cara mentransformasikan tiap-tiap piksel dengan rumus :

1. *Document De-warped*

Untuk meluruskan *text line* yang melengkung, dilakukan dengan cara *fitting* kurva yang cocok dengan kelengkungan teks tersebut. Kelengkungan teks dapat dideteksi dengan cara membuat kurva dari titik tengah tiap-tiap kata seperti yang diilustrasikan pada Gambar 2. Dan perbaikan kelengkungan dapat dilakukan dengan menggunakan rumus :

Dengan *d* adalah perbedaan atau jarak antara *word box* dengan kurva.



**Gambar 2. Kelengkungan text line**



**Gambar 3. Tahap-tahap mengubah permukaan dokumen arab yang melengkung menjadi datar**

Secara umum, tahapan proses yang dikerjakan dalam tugas akhir ini dapat dirangkum dalam Gambar 3.

1. Metodologi

Metodologi yang akan dilakukan dalam Tugas Akhir ini memiliki beberapa tahapan, diantaranya sebagai berikut:

1. Studi Literatur

Pada tahap ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan proses pengolahan citra dokumen terdegradasi, khususnya yang meliputi permasalahan mengenai preprocessing dan segmentation. Literatur yang digunakan dalam pengerjaan Tugas Akhir ini sebagian besar berasal dari internet berupa makalah ilmiah, tesis, artikel, materi kuliah, serta beberapa buku referensi.

1. Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data dengan cara mengambil gambar dokumen arab dengan menggunakan kamera digital. Gambar yang diambil merupakan gambar dokumen yang permukaan halamannya melengkung. Misalnya karena dokumen tersebut mempunyai jumlah halaman yang banyak sehingga bentuknya tebal. Dan jika dibuka akan terdapat halaman yang melengkung dari bagian tengah. Jika halaman dokumen ini dipotret, maka akan menghasilkan gambar seperti pada Gambar 1.

1. Implementasi

Pada tahap ini dilakukan pembuatan system atau program untuk merekonstruksi citra dokumen arab yang terdistorsi. Pembuatan system menggunakan *image processing tool* yaitu Matlab 7.6

1. Pengujian dan Evaluasi

Pada tahap ini dilakukan uji coba terhadap sistem yang telah dibuat, mengamati kinerja sistem yang baru dibuat, serta mengidentifikasi kendala yang mungkin timbul.

1. Penyusunan Buku Tugas Akhir

Tahap terakhir merupakan penyusunan laporan yang memuat dokumentasi mengenai pembuatan serta hasil dari implementasi perangkat lunak yang telah dibuat.

1. Jadwal Kegiatan Tugas Akhir

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Bulan | | | | | | | |
| September | | Oktober | | November | | Desember | |
| 1 | Studi Literatur |  |  |  |  |  |  |  |  |
| 2 | Pengumpulan Data |  |  |  |  |  |  |  |  |
| 3 | Implementasi |  |  |  |  |  |  |  |  |
| 4 | Pengujian dan Evaluasi |  |  |  |  |  |  |  |  |

1. Daftar Pustaka
2. Masalovitch and L. Mestetskiy. "Usage of continuous skeletal image representation for document images de-warping". In 2nd Int. Workshop on Camera- Based Document Analysis and Recognition, Curitiba, Brazil, 2007
3. B.Gatos, I. Pratikakis, and K. Ntirogiannis. "Segmentation based recovery of arbitrarily warped document images". In Proc. Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil, 2007
4. B. Fu, M.Wu, R. Li,W. Li, and Z. Xu. "A model-based book de-warping method using text line detection". In 2nd Int. Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil, 2007
5. Hadi Dehbovid, Farbod Razzazi, Shahpour Alirazee. “ A Novel Method for De-warping in Persian Document Images Captured by Cameras”. International Journal of Image (IJIP), Volume (4): Issue(4). 2010

**LEMBAR PENGESAHAN**

Surabaya, 11 Oktober 2011

Menyetujui,

Pembimbing II

Anny Yuniarti, S.Kom.,M.Comp.Sc

NIP. 198106202005011003

Pembimbing I

Dr. Agus Zainal Arifin, S.Kom, M.Kom

NIP. 197208091995121001