**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

Peradaban Islam pada zaman dahulu telah meninggalkan banyak dokumen-dokumen berharga. Dokumen tersebut perlu dijaga kelestariannya agar bisa terus dikaji isi yang terkandung di dalamnya. Dokumen-dokumen tersebut sebagian besar ditulis dalam bahasa arab. Digitalisasi dokumen arab dirasa sangat diperlukan mengingat banyak dokumen-dokumen arab yang tidak mendapat perwatan yang layak.

Digitalisasi adalah proses penyimpanan dokumen dalam bentuk digital atau *soft copy*, bukan dalam bentuk fisik atau lembaran-lembaran kertas. Jika pembaca ingin membaca buku digital tersebut, maka mereka bisa melakukannya dengan bantuan komputer.

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

Selain dengan menggunakan *scanner*, digitalisasi juga dapat dilakukan dengan menggunakan 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 dokumen, pemindaian dokumen yang tebal juga dapat membuat halaman hasil pemindaian terlihat melengkung. Permukaan halaman yang melengkung ini biasa disebut dengan distorsi geometris.

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. Layaknya perangkat lunak pada umumnya, OCR memiliki keterbatasan yakni tidak dapat mengenali karakter dalam dokumen yang pemukaannya melengkung atau mengalami distorsi geometris.

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*.

Salah satu metode yang berbasis *hardware* 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 sukses 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 tetangga 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 karakter-karakter dalam dokumen arab tersebut dapat dikenali oleh *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

Batasan dalam tugas akhir ini adalah bahwa distorsi geometris merupakan kerusakan yang berupa permukaan halaman yang melengkung.

1. 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

Islam memiliki dokumen-dokumen lama yang perlu dirawat agar tetap terjaga kelestariannya. Perawatan tersebut dapat berupa digitalisasi dokumen.

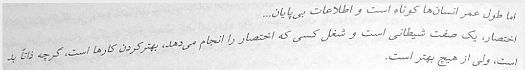
Dokumen dengan halaman yang tebal akan mengalami distorsi geometris jika diakuisisi dengan menggunakan kamera digital atau *scanner*. Distorsi geometris yang dimaksud adalah kerusakan pada permukaan halaman dokumen, misalnya permukaan yang terlihat melengkung. Contoh citra dokumen yang terdistorsi geometris dapat dilihat pada Gambar 1. Citra dokumen yang seperti ini tidak dapat dikenali oleh OCR (*Optical Character Recognition*).

Agar citra dokumen yang mengalami distorsi geometris dapat dikenali OCR, dibutuhkan suatu metode untuk merekonstruksi permukaan dokumen yang melengkung menjadi tidak melengkung. Beberapa metode telah diusulkan, namun sebagian besar metode tersebut dirancang untuk merekonstruksi citra dokumen latin, bukan untuk dokumen arab. Dan jika metode ini diterapkan pada dokumen arab, tidak akan menghasilkan ouput yang berkualitas. Hal ini terjadi karena tulisan arab memiliki karakteristik yang berbeda dengan tulisan latin. Sehingga dibutuhkan suatu metode yang dapat merekonstruksi dokumen arab yang terditorsi geometris agar dokumen tersebut dapat dikenali oleh *Arabic* OCR.

Tugas akhir ini merupakan implementasi metode yang diusulkan oleh Dehbovid, Hadi (2010) [4] yang merupakan pengembangan dari metode *Segmentation Based Recovery* (SEG) [2] untuk merekonstruksi dokumen arab yang terdistorsi geometris. Metode tersebut dapat dijelaskan sebagai berikut :

1. Binarisasi citra

Binarisasi citra adalah mengubah citra berwarna atau citra *grayscale* 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.

Reduksi *noise* dapat dilakukan dalam beberapa tahap, yaitu :

1. Ekstraksi *bounding box*
2. Membuat histogram dari tinggi *bounding box* yang telah didapat sebelumnya
3. Mencari *bounding box* yang mempunyai frekuensi tertinggi, labeli dengan *M*
4. Mengeliminasi *bounding box* yang memiliki tinggi lebih besar dari *a.M*, kurang dari *M.b*, atau 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 *bounding box* pertama pada baris *L* tersebut dilabeli dengan *K*
3. Untuk tiap-tiap baris pada teks, lakukan *right left scanning*
4. *Bounding box* di sebelah kiri *K* yang mempunyai jarak paling dekat dengan *K* akan dilabeli dengan *Kf*. Jika *K* dan *Kf* berjarak kurang dari *threshold* *T*, maka *Kf* dideteksi sebagai tetangga *K* dengan *T* adalah nilai ambang batas jarak yang ditentukan oleh *user*.

Sedangkan untuk deteksi *text line* meliputi deteksi *upper* dan *lower baseline* dari tiap *word box*. Contoh *upper* dan *lower baseline* dari kata dapat dilihat pada Gambar 2. Deteksi *text line* ini dapat dilakukan 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 *word box*, misalnya tinggi *word box*.

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*

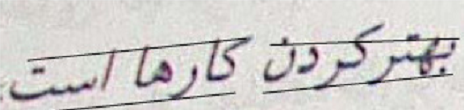
Koreksi *word skew* dapat dilakukan dengan cara mentransformasi tiap-tiap piksel dengan rumus :

Koreksi ini dilakukan berdasarkan kemiringan kata sebelumnya dalam baris yang sama.

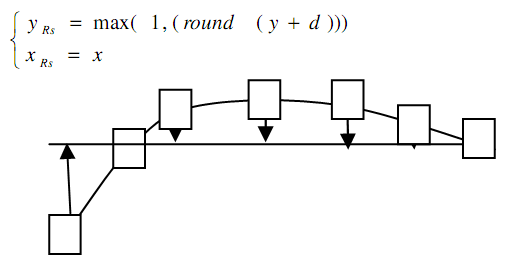
1. *De-warp Document*

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 3. Dan perbaikan kelengkungan dapat dilakukan dengan menggunakan rumus :

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



**Gambar 2. *Upper* dan *lower baseline***



**Gambar 3. Kelengkungan *text line***



**Gambar 4. Tahap-tahap mengubah permukaan dokumen arab yang melengkung menjadi datar**

Secara umum, tahapan proses yang dikerjakan dalam tugas akhir ini dapat dirangkum dalam Gambar 4.

Hasil dari proses rekonstruksi dokumen pada tugas akhir ini dievaluasi dengan cara membandingkan citra dokumen hasil rekonstruksi dengan *ground truth.* Citra hasil rekonstruksi terlebih dahulu diinputkan ke dalam OCR untuk diekstrak karakter-karakter yang terdapat dalam citra tersebut. Kemudian dilakukan perhitungan terhadap jumlah karakter yang dapat dikenali oleh OCR. *Ground truth* diperoleh dengan cara menghitung secara manual jumlah huruf yang terdapat dalam dokumen. Sedangkan akurasi diperoleh dengan cara membandingkan jumlah huruf yang dapat dikenali oleh OCR dengan jumlah huruf pada *ground truth*.

1. Metodologi

Metodologi yang akan dilakukan dalam tugas akhir ini memiliki beberapa tahapan yang dapat dijelaskan sebagai berikut:

1. Studi Literatur

Pada tahap ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan proses pengolahan citra dokumen terdistorsi geometris, khususnya yang meliputi permasalahan mengenai reduksi *noise* dan koreksi kelengkungan teks. Literatur yang digunakan dalam pengerjaan tugas akhir ini sebagian besar berasal dari internet berupa makalah ilmiah, tesis, artikel, materi kuliah, serta beberapa buku referensi. Selain itu, dilakukan juga uji coba pengenalan karakter pada citra dokumen arab yang permukaannya melengkung dengan bantuan *Arabic* OCR. Uji coba ini bertujuan untuk memastikan bahwa karakter-karakter yang terdapat dalam citra dokumen arab yang permukaannya melengkung tidak dapat dikenali oleh *Arabic* OCR.

1. Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data dengan cara mengambil gambar dokumen arab dengan menggunakan kamera digital atau *scanner*. 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 atau dipindai, maka akan menghasilkan gambar seperti pada Gambar 1.

1. Implementasi

Pada tahap ini dilakukan pembuatan program untuk merekonstruksi citra dokumen arab yang terdistorsi geometris. Pembuatan program menggunakan *tools* 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. 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, Mestetskiy. L. "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. 198106222005012002

Pembimbing I

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

NIP. 197208091995121001