**MENGENALI POLA HURUF ARAB MENGGUNAKAN  
PENELUSURAN KONTUR**

Oleh: [Mustafa Ramadhan](https://medium.com/@emramadhan?source=post_page-----2c14cb007453--------------------------------)

*ABSTRAK*

*Tulisan ini menjelaskan mengenai pengenalan huruf arab dengan pendekatan struktural. Setelah akuisisi data, prosesnya adalah segmentasi citra, ekstraksi fitur, dan terakhir klasifikasi huruf. Segmentasi gambar dilakukan dalam dua langkah. Yang pertama adalah segmentasi halaman, dengan hasil segmen garis, yang kedua adalah segmentasi garis, yang menghasilkan segmen kolom huruf. Kolom huruf terdiri dari tiga bagian yaitu bentuk, diakritik*(titik, hamzah حمزة)*, dan tanda vokal ḥarakāt* (حَرَكَات)*. Kami mengekstrak fitur struktural dari ketiga bagian ini dengan menelusuri gambar kontur. Klasifikasi huruf didasarkan pada hasil ekstraksi ciri bentuk dan diakritik. Klasifikasi vokal didasarkan pada hasil ekstraksi ciri tanda vokal. Dari klasifikasi huruf dan tanda vokal diperoleh teks Arab dalam rich text format (RTF) dan disimpan dalam file. File dapat dibuka dan diedit menggunakan pengolah kata seperti Wordpad atau Word.*

**Pendahuluan**

Tiga atau empat dekade terakhir, banyak upaya telah dilakukan untuk mengenali huruf Latin. Sistem pengenalan huruf latin optik telah berhasil dikembangkan. Sistem komersial tersedia dan digunakan secara luas. Sebaliknya, upaya pengenalan huruf Arab masih terbatas. Beberapa di antaranya adalah sebagai berikut.

Hussain dan Cowell (2000:2–3) mengusulkan pendekatan pencocokan templat untuk mengenali huruf Arab yang terisolasi. Mereka menggunakan 100 x 100 piksel sebagai template referensi. Akibatnya timbul kerancuan dalam mengenali huruf serupa. Huruf yang paling mirip adalah 96% غ ع %88 ,ظ ط %93 ,ج ح.

Proses penipisan huruf merupakan langkah awal yang umum dilakukan untuk mendapatkan ciri-ciri geometris. Fahmy (2001:7–14) menggunakan pendekatan ini untuk mengenali huruf Arab tulisan tangan. Berdasarkan fitur geometris, klasifikasi huruf dilakukan menggunakan jaringan saraf tiruan propagasi mundur kesalahan feed-forward.

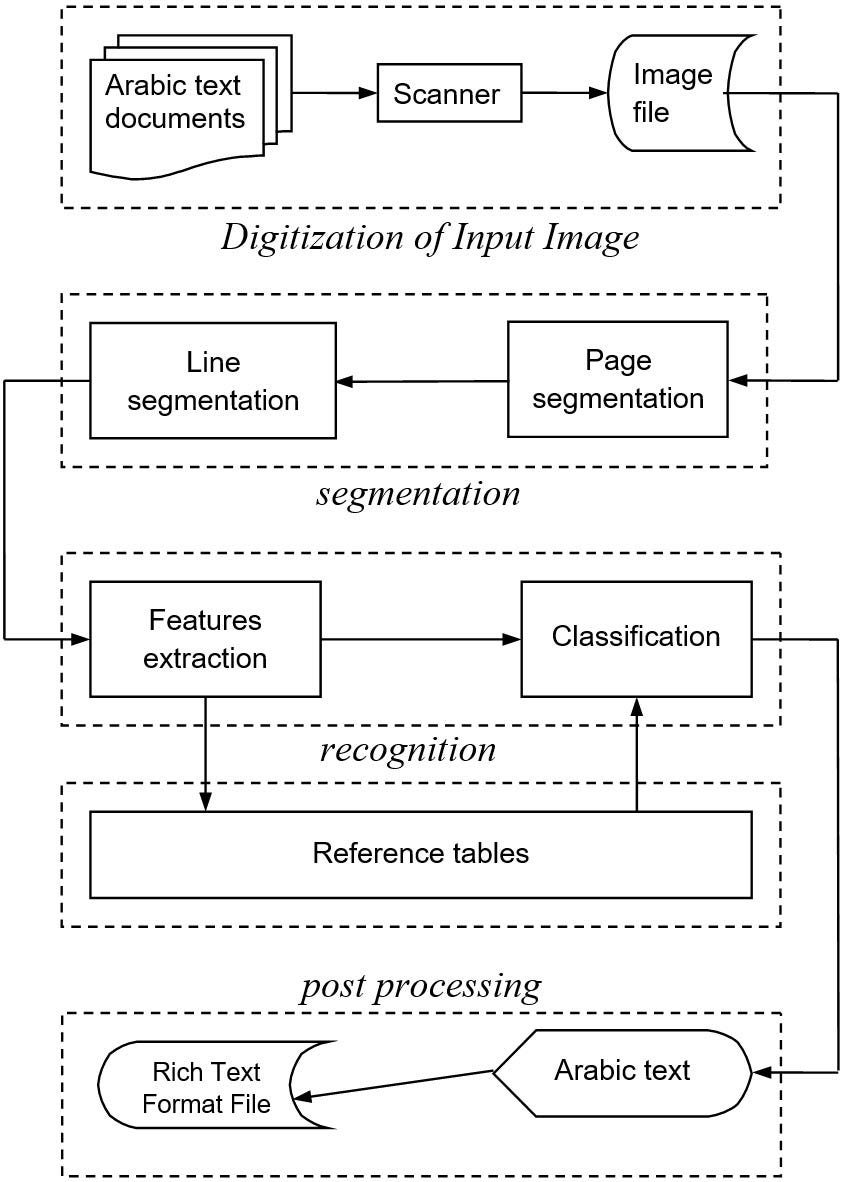
Mitiche (2001:5–8) memperkenalkan sistem pengenalan untuk huruf cetak Arab multi-font dan multi-ukuran. Pengklasifikasian huruf dilakukan dengan menggunakan ciri-ciri huruf seperti jumlah lubang, jumlah cekungan dan arah, jumlah titik, dan lain-lain.

Bahasa Arab digunakan di lima belas negara Arab di Timur Tengah dan Afrika Utara sebagai bahasa lisan. Huruf Arab banyak digunakan oleh orang Arab dan non-Arab seperti Kurdi, Persia, Urdu, dan Turki sebagai bahasa tulisan. Di negara lain seperti Afghanistan, Kazakhstan, Turkmenistan, Bangladesh, Malaysia, Brunei, Indonesia, dll., teks Arab juga dikenali dan digunakan. Sekitar satu miliar orang yang menggunakannya. Sehingga tulisan ini dapat memberikan kontribusi bagi pengembangan sistem pengenalan huruf Arab. Suatu hari nanti, sistem pengenalan huruf Arab yang andal tersedia secara luas dan digunakan oleh orang Arab dan non-Arab yang menggunakan huruf Arab sebagai bahasa tertulis.

**Pengenalan Huruf Arab**

Teknik yang berhasil pada huruf Latin tidak dapat diterapkan dengan tingkat keberhasilan yang sama pada huruf Arab. Sistem pengenalan huruf latin, misalnya, tidak mempertimbangkan jumlah atau posisi titik, yang merupakan hal krusial dalam huruf Arab. Karena tulisan Arab mempunyai ciri-ciri khusus, maka pendekatan yang digunakan untuk huruf Latin tidak dapat dialihkan secara langsung.

Makalah ini menjelaskan pengenalan huruf pada dokumen cetak berbahasa Arab melalui lima langkah, yaitu (1) digitalisasi citra masukan, (2) segmentasi, (3) ekstraksi fitur, (4) pengambilan keputusan dan klasifikasi, dan (5) pasca pemrosesan. . Konstruksi sistem pengenalan huruf arab ditunjukkan pada Gambar 1.



Gambar 1 Konstruksi sistem pengenalan huruf arab

**Ciri-ciri Huruf Arab**

Huruf arab mempunyai ciri-ciri utama sebagai berikut:

* Teks berbentuk kursif dari kanan ke kiri.
* Huruf arab terdiri dari 28 huruf dasar dan 10 huruf tambahan.

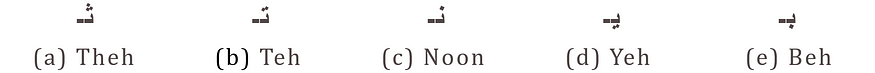
**[Alfabet Arab](http://mylanguages.org/arabic_alphabet.php?source=post_page-----2c14cb007453--------------------------------" \t "_blank)**

[Jika Anda mencoba mempelajari Alfabet Arab, Anda akan menemukan beberapa sumber berguna, termasuk kursus tentang…](http://mylanguages.org/arabic_alphabet.php?source=post_page-----2c14cb007453--------------------------------" \t "_blank)

[bahasaku.org](http://mylanguages.org/arabic_alphabet.php?source=post_page-----2c14cb007453--------------------------------" \t "_blank)

Beberapa huruf terletak di bawah garis pangkal, misalnya و ز ر.

* Huruf Arab yang berbeda mungkin mempunyai bentuk yang sama, dan hanya dapat dibedakan satu sama lain dengan menambahkan beberapa titik dan posisinya (Gambar 2).



Gambar 2 Huruf Arab yang berbeda berdasarkan  
jumlah titik yang terkait dan posisinya

* Berdasarkan bentuknya, seluruh huruf dapat dikelompokkan menjadi 18 kelompok (Tabel 1).

Tabel 1 Kelompok Huruf Arab Berdasarkan Kesamaan Bentuk



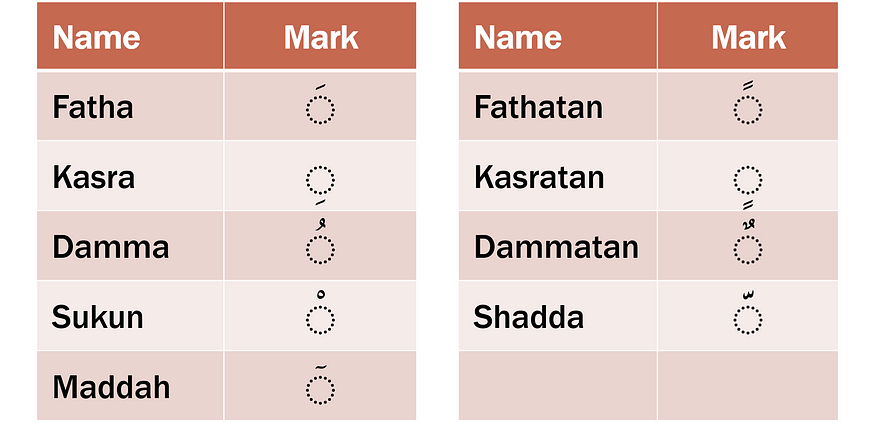
* Kebanyakan huruf Arab memiliki hingga empat bentuk berbeda. Tergantung posisinya: di awal kata (Gambar 3a), di tengah kata (Gambar 3b), di akhir kata (Gambar 3c), atau berdiri sendiri (Gambar 3d). Lebar dan panjang huruf berbeda antara satu huruf dengan huruf lainnya.



Gambar 3 Bentuk *Ghain yang berbeda*

* Dalam penyajian huruf vokal, bahasa Arab menggunakan tanda vokal *ḥarakāt* (حَرَكَات) (Tabel 2).

Tabel 2 Harakat



**Digitalisasi Gambar Masukan**

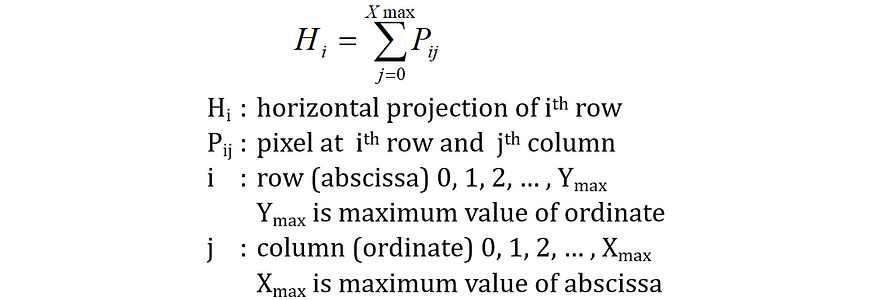
Inputnya berupa gambar biner yang diperoleh dengan memindai suatu dokumen dengan resolusi yang sebanding dengan ukuran font yang digunakan pada dokumen tersebut.

**Segmentasi**

Gambar teks diubah menjadi matriks biner nol dan satu. Segmentasi citra dilakukan dalam dua tahap: (1) segmentasi halaman menghasilkan segmen garis, dan (2) segmentasi garis menghasilkan segmen kolom huruf.

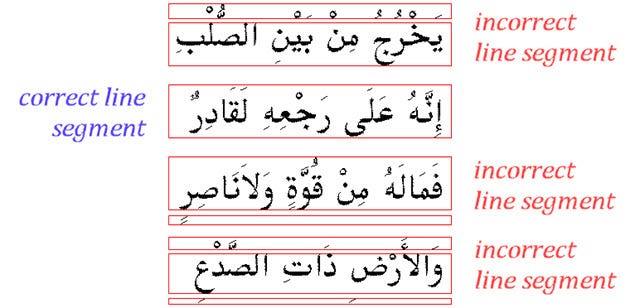
**Segmentasi Halaman**

Teks dipartisi menjadi garis-garis penyusunnya, menggunakan proyeksi horizontal gambar yang ditemukan dengan menghitung piksel di setiap baris matriks. Proyeksi horizontal baris ke-i adalah:



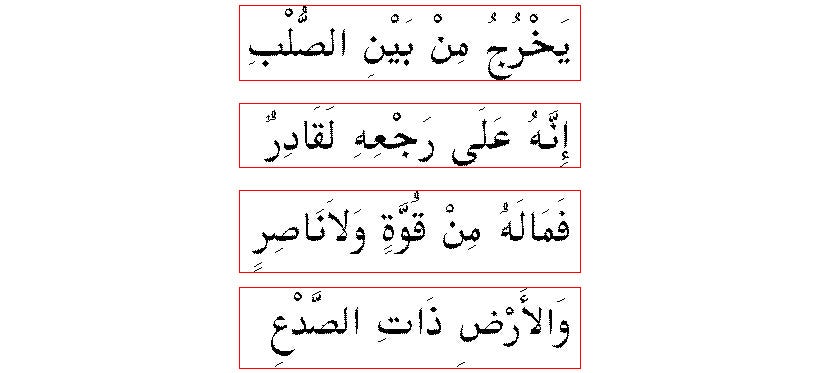
Proyeksi horizontal Formula 1

Ada dua langkah dalam proses segmentasi. Langkah pertama mungkin menghasilkan segmen garis yang salah (Gambar 4).



Gambar 4 Hasil langkah pertama segmentasi halaman

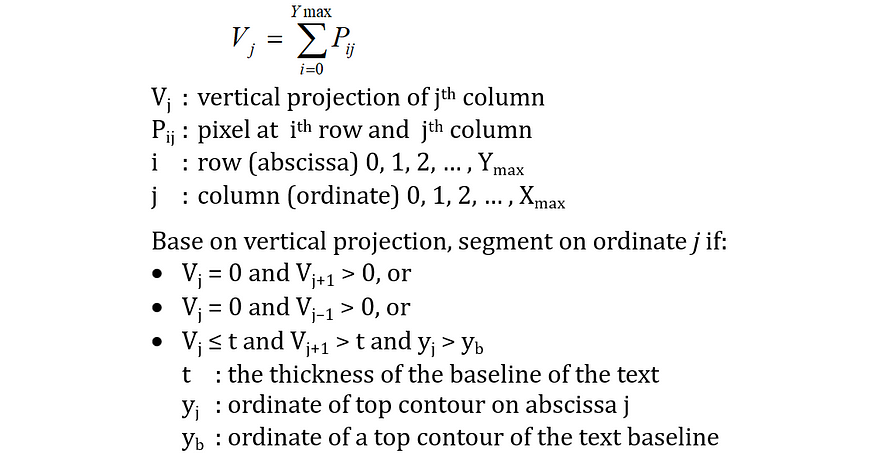
Langkah kedua untuk memperbaiki ruas garis yang salah adalah dengan membandingkan tinggi ruas garis tersebut. Hasilnya, semua ruas garis benar (Gambar 5).



Gambar 5 Hasil segmentasi halaman langkah kedua

**Segmentasi Garis**

Segmen garis gambar teks dipartisi menjadi huruf-huruf penyusunnya, menggunakan proyeksi vertikal gambar yang ditemukan dengan menghitung piksel di setiap kolom matriks. Proyeksi vertikal baris ke-i adalah:



Proyeksi Vertikal Formula 2

Mengolah garis satu per satu, ditemukan proyeksi vertikal setiap garis. Nilai minimum proyeksi vertikal menentukan titik di mana teks dipecah menjadi huruf-huruf (Gambar 6). Namun tidak semua huruf bisa dipisahkan dengan cara ini. Perlakuan khusus pada tahap selanjutnya diperlukan untuk memisahkan huruf-huruf apabila terdapat tumpang tindih antar huruf akibat adanya perpanjangan huruf di atas/di bawah huruf yang lain.



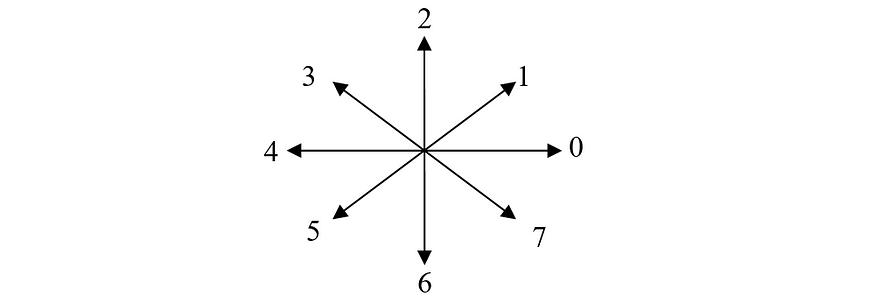
Gambar 6 Segmen huruf

**Fitur Ekstraksi**

Kolom Surat, yaitu sebuah huruf yang terdiri atas tiga bagian, yaitu (1) bentuk, (2) diakritik (titik, *hamzah* ), dan (3) tanda vokal *ḥarakāt* . Kami mengekstrak fitur struktural dari ketiga bagian ini dengan menelusuri kontur gambar. Proses penipisan gambar huruf tidak diperlukan. Algoritma yang ada saat ini untuk komputasi kerangka pada teks kursif masih buruk (Steinherz, 2000:529). Apalagi proses penipisan tidak cocok untuk huruf sambung seperti teks arab. Ada beberapa permasalahan dalam penerapannya (Hussain, 2001:2–4).

**Ekstraksi Fitur Bentuk**

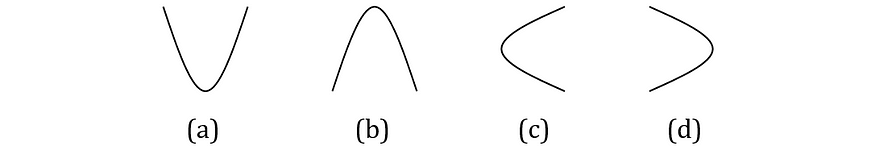
Penelusuran gambar kontur menggunakan vektor delapan arah (Gambar 7). Ini digunakan untuk mencatat arah penelusuran.



Gambar 7 Vektor delapan arah

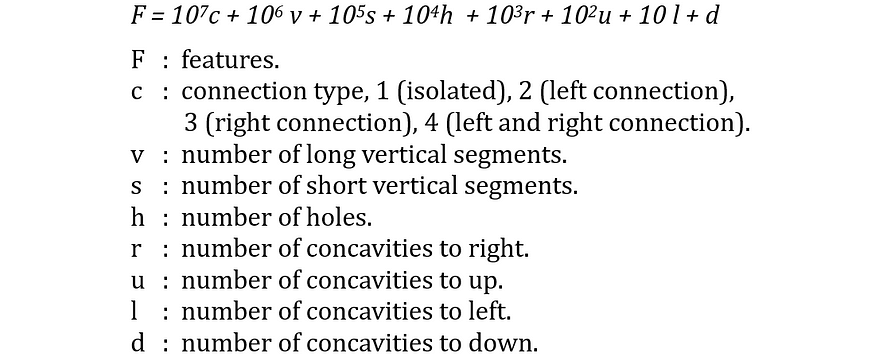
Ada lima jenis fitur yang perlu diekstraksi, yaitu:

* Jenis sambungan: sambungan terisolasi (berdiri sendiri), kiri, kanan, atau kiri dan kanan.
* Jumlah segmen vertikal yang panjang.
* Jumlah segmen vertikal pendek.
* Jumlah lubang.
* Banyaknya cekungan ke kiri, kanan, atas, dan bawah.



Gambar 8 Cekungan ke (a) atas, (b) bawah, © kanan, dan (d) kiri.

Semua fitur setiap huruf disimpan sebagai nilai integer mengikuti rumus:



Rumus 3 Ciri-ciri surat

**Ekstraksi Fitur Diakritik**

Ciri-ciri yang perlu digali adalah:

● Jenis diakritik (titik, *hamzah* )  
● Posisi dan jumlah titik  
● Posisi *hamzah* .

**Ekstraksi Fitur Ḥarakāt**

Ada sembilan *ḥarakāt* (Tabel 2). Ciri-ciri yang perlu digali adalah:  
● jenis *ḥarakāt*● posisi relatif terhadap garis dasar  
● jumlah *ḥarakāt* .

**Klasifikasi**

Klasifikasi didasarkan pada hasil ekstraksi fitur pada langkah sebelumnya. Jadi, ada tiga jenis klasifikasi yaitu  
(1) klasifikasi bentuk,  
(2) klasifikasi huruf, dan  
(3) klasifikasi *ḥarakāt .*

**Klasifikasi Bentuk**

Huruf Arabnya ada 38 huruf, yakni 28 huruf dasar dan 10 huruf tambahan. Semua huruf tersebut dapat diklasifikasikan menjadi 18 kelompok (Tabel 1). Fitur yang diekstraksi pada langkah sebelumnya, nilai integer yang dihitung dari rumus (3), memberikan informasi yang diperlukan untuk mengidentifikasi bentuk huruf.

**Klasifikasi Surat**

Identifikasi huruf diperoleh dari hasil klasifikasi bentuk yang digabungkan dengan hasil ekstraksi ciri diakritik pada langkah sebelumnya.

**Klasifikasi Ḥarakāt**

Identifikasi ḥarakāt *diperoleh* dari hasil ekstraksi ciri *ḥarakāt* pada langkah sebelumnya.

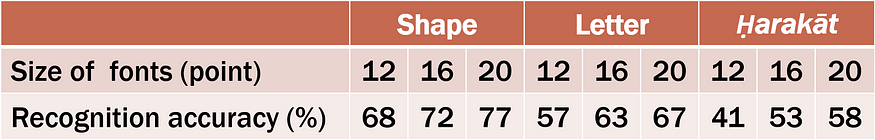
**Pengolahan pasca**

Hasil klasifikasi huruf dan klasifikasi *ḥarakāt* disimpan dalam file rich text format (RTF). File dapat dibuka dan diedit menggunakan pengolah kata seperti Wordpad atau Word.

**Hasil**

Sistem diuji pada pemindaian dokumen berbahasa Arab resolusi 300x300 menggunakan font *Akhbar MT*12, 16, dan 20 poin . Hasil pengujian pada mesin dengan prosesor Duron (900 MHz), RAM 128 MB, ditunjukkan pada Tabel 3 dan Tabel 4.

Tabel 3 Keakuratan pengenalan bentuk, huruf, dan *ḥarakāt*



Tabel 4 Kecepatan pengenalan



Huruf yang kurang dikenali secara akurat ditunjukkan pada Tabel 5. Akurasi rendah karena pengenalan diakritik (titik, *hamzah* ) tidak akurat. Apalagi sulit membedakan bentuk *ḥarakāt* kecil . Hasilnya akan lebih baik jika dilakukan preprocessing sebelum melakukan ekstraksi fitur. Dengan prapemrosesan, kami mengacu pada semua tugas yang diperlukan untuk menyiapkan gambar untuk ekstraksi fitur. Gambar teks dibersihkan dari noise dengan menghilangkan detail yang tidak relevan.

Tabel 5 Pasangan huruf yang mungkin tidak dikenali secara akurat



**Kesimpulan**

Metode pengenalan huruf Arab pada dokumen hasil cetakan mesin dikembangkan berdasarkan fitur struktural dengan menelusuri gambar kontur. Kinerja sistem memadai, namun keakuratan pengenalannya masih buruk. Akurasinya dapat ditingkatkan dengan melakukan pra-pemrosesan untuk mengurangi noise, sehingga ekstraksi fitur hanya menggunakan detail yang relevan.

**DAFTAR PUSTAKA**

Brondgeest, BT (1926). *Moechtasar 'Ilmoe Bahasa 'Arab.*Landsdrukkerij-Weltevreden: Balai Poestaka.

Fahmy, Maged Mohamed Mahmoud. (2001). “Pengenalan Otomatis Karakter Arab Tulisan Tangan Menggunakan Ciri Geometrisnya”. Bahrain: Sekolah Tinggi Sains Departemen Ilmu Komputer Universitas Bahrain.

Hussain, Fiaz & Cowell, John. (2000). “Algoritma Pengenalan Generik untuk Aksara Latin dan Arab”. Dubai: Sekolah Teknologi Informasi.

Hussain, Fiaz & Cowell, John. (2001). “Menipiskan Karakter Arab untuk Ekstraksi Ciri”. Dubai: Sekolah Teknologi Informasi.

Mitiche, L. Hamami. (2000). “Sistem Pengenalan Karakter Tulisan Arab Multi Font dan Multi Ukuran”. *Prosiding Kongres Dunia IMACS ke-16* .

Schalkoff, Robert J. (1992). *Pengenalan Pola: Pendekatan Statistik, Struktural dan Neural* . New York: John Wiley & Sons, Inc.

Steinherz, Tal, dan Nathan Intrator. (2000). “Algoritma Skeletonisasi Khusus untuk Kata Kursif”, *Prosiding Lokakarya Internasional ke-7 tentang Frontiers dalam Pengenalan Tulisan Tangan* . Amsterdam: 11–13 September 2000.