Pengecaman Tulisan Tangan: Keperluan Sistem dan Satu Pendekatan Terhadap penyelesaian Deterministik

Musa Md. Lazim
Mohd. Noor Md. Sap
Dzulkifli Mohamad

Institut Sains Komputer
Universiti Teknologi Malaysia

Abstrak

Di dalam kertas-kerja ini, keperluan-keperluan minima terhadap perkakasan dan perisian dan juga ciri-ciri penting bagi sesuatu sistem pengecaman teks tulisan tangan yang hendak dibangunkan adalah dibincangkan. Kejayaan sesuatu sistem pengecaman teks atau aksara beroptik (OCR - Optical Character Recognition) adalah bergantung kepada sejauh mana sesuatu sistem itu dapat mengecami apa jenis bentuk penulisan - tulisan yang tidak mempunyai kekangan (sebarang bentuk sama ada besar, kecil, skrip, dan lain-lain) dan mempunyai peratus kesalahan pengecaman yang sangat kecil. Keupayaan atau kebolehan ini sangat bergantung kepada keberkesanan metodologi yang dibangunkan, terutamanya di dalam kaedah pelicinan dan pengecilan imej bagi teks yang diimbas - proses pra-pemorosesan. Ketepatan pengecaman boleh dibuat, jika sifat-sifat imej yang dikumpul(diambil dari pengajian terhadap sifat topologi, goemetri dan pengiraan-pengiraan tertentu) dapat dikenali dengan baik untuk dipadankan dengan sifat-sifat alphabet di dalam pengkalan data.

Abstract

In this paper, the minimum requirements of hardwares, sofware: and the important characteristics of any system of tex recognition to be developed is described. The success of an) texts recognition or optical character recognition system i: merely depended on the ability to recognise any type o unconstrained characters, and the ability to minimise the recognition error. This ability is also depended on the effectiveness of methodology used, especially in the process o) smoothing and thinning of the scanned image during pre-processing phase. The correctness of recognition can be achieve(if the extracted features(i.e. from the topological, geometry an(analysis properties of image)can be properly extracted and matcl to the stored features in the data base.

Katakunci : *Pengecaman teks, pengecaman Corak, Pengumpulan Sifat-sifat, imej raster, Aksara beroptik, geraf-bit, Pelicinan imej, Pengecilan imej.*

1. Pengenalan

Pengecaman teks atau aksara adalah merupakan subset dari bidang pengecaman corak. Sebelum kita pergi lebih jauh lagi, eloklah kita memandang sepintas lalu tentang apakah itu pengecaman corak. Masalah tentang bidang ini selalunya menerangkan satu diskriminasi atau pemilihan suatu set proses-proses atau peristiwa-peristiwa.

Set proses-proses atau peristiwa-peristiwa yang hendak dipilih ini, boleh jadi satu set objek-objek fisikal atau satu set keadaan-keadaan hasil dari sesuatu penilaian yang dibuat. Jumlah kelas corak-corak biasanya ditentukan mengikut keadaan di mana ia hendak digunakan. Secara keseluruhannya, pengecaman corak ini boleh dipandang sebagai satu 'perkakas' yang membantu di dalam membuat kata-putus automatik; yang memindahkan pengiraan terhadap corak kepada kelas-kelas tertentu. Corak-corak ini sendiri boleh jadi digunakan di dalam bidang pertanian yang diambil dari imej pengalamatan jauh melalui satelit Landsat misalnya, sehinggalah kepada bentuk gelombang hasil dari pertuturan seseorang.

Pengecaman terhadap masalah-masalah tersebut dikelaskan mengikut bidang keperluan masing-masing; misalnya dibidang pertanian tadi kita ingin menentukan sama ada corak tadi menunjukan kita kepada tanaman padi atau getah dan di dalam bidang pengecaman suara pula kita ingin mengenali apakah yang diucapkan oleh seseorang itu. Corak-corak tersebut dicam berpandukan kepada sifat-sifat, ciri-ciri atau pengiraan-pengiraan tertentu yang dibuat ke atas corak tersebut. Di dalam penggunaan pengalamatan jauh, sifat-sifat yang hendak dikaji adalah berupa tenaga di dalam beberapa panjang-gelombang spektrum elektromagnet yang dibalekkan; manakala didalam masalah pengecaman suara, set bagi sifat-sifat yang diselidiki diambil dari angkali ramalan linear yang dihasilkan dari bentuk gelombang percakapan. Bagi masalah pengecaman teks pula, sifat-sifat ini diambil dari bentuk geometriknya seperti garis lengkung, bulatan, garis hujung atau garis silang.

2. Latarbelakang Sejarah

Minat di dalam pengecaman teks ini, telah lama bermula, seawal pertengahan tahun 1940; sejajar dengan perkembangan komputer digit. Pada akhir-akhir ini, pengecaman aksara beroptik (OCR - Optical Character Recognition) iaitu sistem pengecaman terhadap tulisan yang dihasilkan oleh mesin taip, telah menjadi satu hakikat perdagangan. Seperti pemerosesan data yang lain, ianya telah digunakan secara meluas di dalam dunia perniagaan dengan tujuan-tujuan tertentu.

Dimasa sekarang fokus utama bidang ini adalah menuju kearah bagaimana dapat diperluaskan lagi teknik pengecaman bagi meliputi pelbagai jenis tulisan

teks yang tiada mempunyai kekangan. Kekangan-kekangan yang ada sekarang dan perlu diatasi termasuklah;

i. Tulisan-tulisan tangan yang tidak bebas - Teknik yang ada sekarang tidak cekap dalam mengecam tulisan-tulisan yang ditulis secara bersambung dan 'berbunga'. Penekanan hanya dibuat keatas tulisan-tulisan berbentuk tunggal.

ii. Tulisan-tulisan mesin taip yang terhad - Sistem pengecaman aksara beroptik yang sedia ada sekarang tidak mampu mengecam semua jenis huruf mesin taip.

iii. Tulisan-tulisan bagi bahasa yang tidak menyeluruh - Pengecaman teks yang dibangunkan sekarang tidak meliputi kesemua bahasa yang terdapat didalam dunia ini. Setakat ini, bahasa-bahasa seperti Arab, Kanji, Cina, Telegu, Hebrew dan Inggeris telah pun dipilih oleh penyelidik-penyelidik diserata dunia untuk dilakukan proses pengecaman keatasnya.

3. Di manakah Diperlukan Pengecaman Teks?

Sebelum ini, penggunaan pengecaman teks tidak begitu ketara dan hanya merupakan aktibiti-aktibiti biasa di dalam masalah pengecaman corak. Sebaliknya hari ini, oleh kerana keperluan yang begitu besar dan mendadak bagi teknologi pemerosesan maklumat di dalam mengoptimumkan pengkomputeran pejabat, maka penggunaan pengecaman teks amatlah dirasai. Oleh kerana kaedah pemerosesan dan penyimpanan maklumat masih dan dijangka kekal dengan cara membaca dan mencetak diatas kertas, maka satu teknik bagi mengecam teks dari bentuk asalnya ke dalam bentuk yang boleh diproses oleh komputer secara automatik adalah diperlukan.

Penggunaan pengecaman teks automatik dimasa sekarang telah digunakan secara meluas di dalam tiga persekitaran berikut;

i. Kemasukan data didalam persekitaran institusi bank.

Didalam persekitaran bank, pengecaman teks telah digunakan untuk mengecam sejumlah set aksara (nombor dan simbol khas) yang ditulis diatas cek atau kertas. Ini adalah untuk mendapatkan pengawasan yang kukuh terhadap kesalahan didalam membaca dan mencetak.

ii. Kemasukan teks didalam persekitaran pejabat.

Didalam persekitaran pengkomputeran pejabat pula, pengecaman teks digunakan untuk mengecam sesuatu dokumen didalam format kertas biasa samaada ianya adalah di dalam tulisan mesin taip atau tulisan tangan dan kebiasaannya pengguna bekerja didalam suasana pemerosesan perkataan.

iii. Proses pengkomputeran dipersekitaran pejabat pos.

Manakala didalam persekitaran pejabat pos pula, penggunaanya adalah lebih kepada pengecaman automatik terhadap alamat pos untuk maksud melakukan proses penyusunan dan klasifikasi alamat-alamat yang tertulis di atas sampul-sampul surat.

4. Konfigurasi Bagi Sistem Pengecaman Teks

Pembangunan sesuatu sistem pengecaman teks, pada keseluruhannya akan membabitkan tiga fasa utama: pengambilan data, pra-pemerosesan data, dan klasifikasi kata-putus, seperti yang ditunjukan di dalam gambarajah 1.

4.1 Fasa pengambilan data

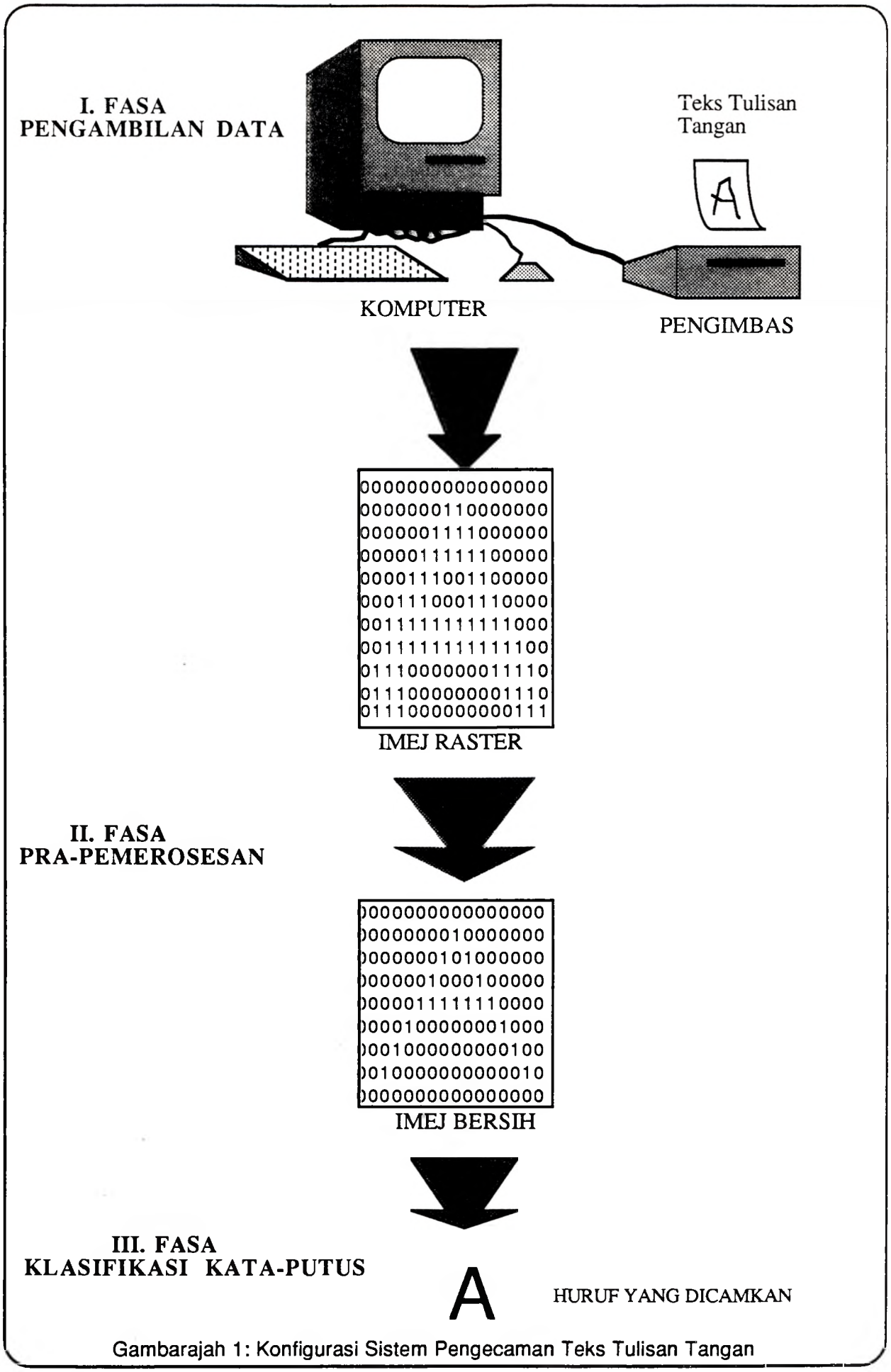
Teks atau data berbentuk analog, didalam sebutan lain yang lebih umum; diimbas dari dunia fisikal dan ditukar kepada bentuk signal binari - 1 dan 0 (imej raster) - untuk pemerosesan elektronik selanjutnya. Oleh kerana penulisan teks selalunya hanya melibatkan satu jenis warna (dakwat), maka pengimbas tidak akan menaksirkan digit 1 dan 0 ke dalam beberapa paras kelabu seperti gambar atau grafik yang kompleks. Setelah pengimbas menukarkan teks asal ke dalam bentuk geraf-bit (bit-mapped), maka langkah selanjutnya adalah menyimpan imej ke dalam RAM komputer atau cakera keras sebagai input bagi sistem pengecaman yang akan dibuat..

Sistem pengimbasan yang sangat popular hari ini adalah terdiri dari satu pengimbas optik yang dapat menjana dan menghasilkan satu imej hitam dan putih bagi grafik atau gambar yang diimbas. Muka surat yang berukuran piawai (A4 - 8.5 X 11 inci) boleh diimbas pada kadar 300 titik-titik bagi setiap inci dan dapat menjana sehingga 1 Megabait data imej. Satu lagi pendekatan di dalam pengecaman teks atau aksara ini, adalah dikenali sebagai pengecaman aksara bertalian terus; di mana aksara dilukis di atas tablet grafik dan kemudiannya dicamkan terus oleh komputer. Walau bagaimanapun, sistem pengecaman talian terus ini agak ringkas dan lebih mudah jika dibandingkan dengan sistem pengimbasan yang menggunakan alat pengimbas.

4.2 Fasa pengambilan data

Teks atau data berbentuk analog, didalam sebutan lain yang lebih umum; diimbas dari dunia fisikal dan ditukar kepada bentuk signal binari - 1 dan 0 (imej raster) - untuk pemerosesan elektronik selanjutnya. Oleh kerana penulisan teks selalunya hanya melibatkan satu jenis warna (dakwat), maka pengimbas tidak akan menaksirkan digit 1 dan 0 ke dalam beberapa paras kelabu seperti gambar atau grafik yang kompleks. Setelah pengimbas menukarkan teks asal ke dalam bentuk geraf-bit (bit-mapped), maka langkah selanjutnya adalah menyimpan imej ke dalam RAM komputer atau cakera keras sebagai input bagi sistem pengecaman yang akan dibuat..

Sistem pengimbasan yang sangat popular hari ini adalah terdiri dari satu pengimbas optik yang dapat menjana dan menghasilkan satu imej hitam dan putih bagi grafik atau gambar yang diimbas. Muka surat yang berukuran piawai (A4 - 8.5 X 11 inci) boleh diimbas pada kadar 300 titik-titik bagi setiap inci dan dapat menjana sehingga 1 Megabait data imej. Satu lagi pendekatan di dalam pengecaman teks atau aksara ini, adalah dikenali sebagai pengecaman aksara bertalian terus; di mana aksara dilukis di atas tablet grafik dan kemudiannya dicamkan terus oleh komputer. Walau bagaimanapun, sistem pengecaman talian terus ini agak ringkas dan lebih mudah jika dibandingkan dengan sistem pengimbasan yang menggunakan alat pengimbas.



Gambarajah 1: Konfigurasi Sistem Pengecaman Teks Tulisan Tangan

4.3 Fasa pra-pemerosesan

Tujuan fasa ini adalah untuk memproses imej yang berada di dalam RAM atau cakera keras tadi bagi menghasilkan imej yang berada di dalam keadaan 'bersih' atau di dalam keadaan 'bersedia' untuk dicamkan didalam fasa ketiga nanti. Secara keseluruhannya fasa ini akan melibatkan dua proses yang utama.

a. Proses pengecilan imej raster kepada imej yang terdiri dari titik-titik linear yang tunggal.
b. Proses pengumpulan sifat-sifat/ciri-ciri geometri atau penentuan bentuk imej.

Terdapat pertalian yang sangat rapat dan penting di antara kedua-dua proses tersebut. Penentuan sifat-sifat atau bentuk ini akan gagal atau disalah ertikan jika imej tidak langsung atau terlalu sedikit mengalami proses pra-pemerosesan. Jika hal ini berlaku, maka ini akan menyebabkan berlakunya penyelewangan atau kesalahan di peringkat klasifikasi kata-putus pengecaman di dalam fasa ketiga.

4.3.1 Pelicinan dan Pengecilan Imej

Maksud dari pengecilan imej raster ialah mengurangkan piksel-piksel hitam yang terdapat diimej asal sehingga meninggalkan seakan-akan satu garisan tengah yang dihubungkan oleh piksel-piksel hitam yang tunggal (sila lihat dalam gambarajah 1, bagi fasa pra-pemerosesan). Terdapat beberapa masalah 'gangguan' berlaku apabila sesuatu imej itu diimbas. Jenis-jenis gangguan yang dimaksudkan itu adalah seperti berikut; (seperti yang ditunjukkan di dalam gambarajah 2)

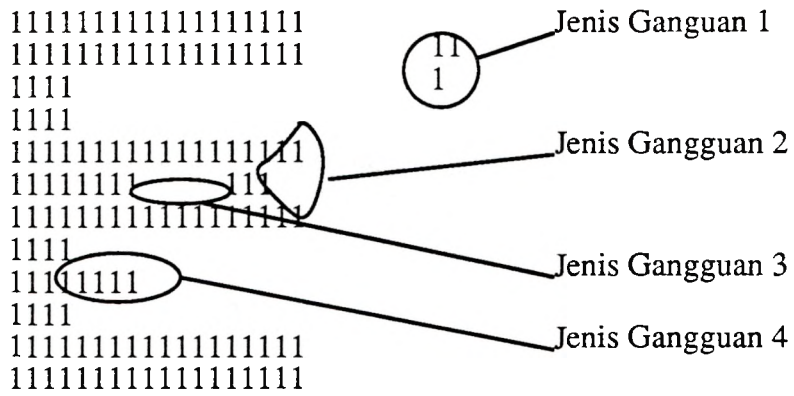
1. Titik-asing: Ini terhasil disebabkan sama ada terdapat kekotoran pada kertas yang diimbas atau terdapat serbuk pada aksara lain yang melekat pada pengimbas.

2. Tepi yang kasar: Sempadan luar/tepi seperti ini terhasil akibat ketidaksamaan terhadap kontur aksara. Dan ini pula adalah disebabkan oleh penulisan teks yang 'kasar'.

3. Lubang buatan: Lubang ini terjadi disebabkan 1 atau beberapa piksel hilang.

4. Garisan buatan: Garisan-garisan ini mungkin panjang atau pendek yang terhasil kadang-kadang akibat kesan dari pengimbas.

Oleh itu, sebelum proses pengecilan dapat dilakukan ke atas imej binari, maka kita harus mengambil kira terhadap kemungkinan berlakunya gangguan-gangguan tersebut, dan jika ini ujud maka ia hendaklah dibersihkan terlebih dahulu melalui proses pelicinan. Beberapa algorithma telah dibangunkan (Smith[1987], Chu[1986]) untuk tujuan ini dan boleh digunakan. Jika algorithma lain yang difikirkan lebih efektif hendak dibangunkan maka haruslah mengambil kira terhadap jenis-jenis gangguan yang telah dibincangkan



Gambarajah 2: Jenis-jenis Gangguan

Oleh itu, sebelum proses pengecilan dapat dilakukan ke atas imej binari, maka kita harus mengambil kira terhadap kemungkinan berlakunya gangguan-gangguan tersebut, dan jika ini ujud maka ia hendaklah dibersihkan terlebih dahulu melalui proses pelicinan. Beberapa algorithma telah dibangunkan (Smith[1987], Chu[1986]) untuk tujuan ini dan boleh digunakan. Jika algorithma lain yang difikirkan lebih efektif hendak dibangunkan maka haruslah mengambil kira terhadap jenis-jenis gangguan yang telah dibincangkan.

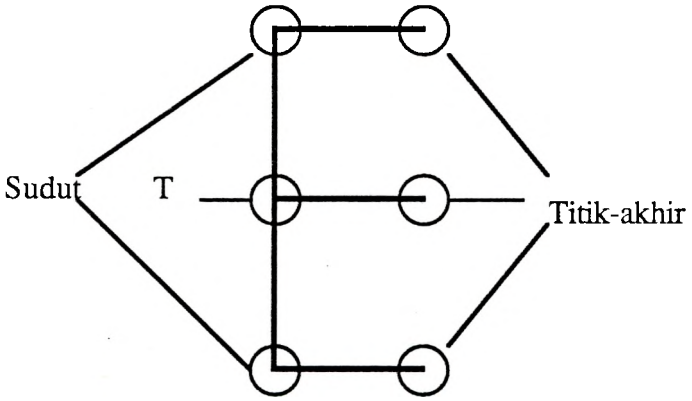
4.3.2 Penakrifan dan Pengumpulan Sifat-sifat

Penakrifan suatu set bagi sifat-sifat aksara/simbol yang terkandung di dalam imej yang diimbas adalah merupakan perkara pokok di dalam setiap sistem pengecaman yang hendak dibangunkan. Cara bagaimana sifat-sifat itu dipilih dan dikumpul akan memberi kesan yang ketara bagi pengecaman yang akan diadili kelak. Misalnya, huruf 'D' berkemungkinan besar dicamkan sebagai huruf 'O', kerana sifatnya saling menyerupai; iaitu satu lubang ditengah. Huruf I mungkin dicamkan sebagai huruf J, sebab sifat bagi hujungnya seakan-akan serupa.

Beberapa pengkelasan sifat-sifat diberikan di sini untuk tujuan pertimbangan terhadap bagaimana ciri-ciri istimewa itu diambil daripada calun imej untuk dipadankan dengan huruf sasaran di dalam klasifikasi kata-putus.

Sifat-sifat Diperingkat Rendah - Sifat-sifat pada peringkat ini sangat ringkas dan mudah ditakrifkan, akan tetapi menyumbangkan maklumat yang sangat penting kepada sebarang algorithma pengecaman tulisan tangan yanghendak dibangunkan. Keujudan *titik-akhir* (penghujung terakhir bagi sesuatu garisan) dan *titik-T* (bila terdapat dua garis bersilang di satu penghujung.) memberikan satu ciri istimewa yang berguna. Setelah sifat-sifat ini ditemui, maka kedudukan relatif di antara titik-akhir/titik-T terhadap aksara secara keseluruhannya digunakan untuk menentukan subset bagi set ciri-ciri sesuatu aksara. Contohnya, huruf 'E' (lihat di dalam gambarajah 3) mempunyai 3 - titik-akhir, 1 - titik-T dan 2 - sudut merupakan 3 subset penting bagi set ciri-ciri terhadap keseluruhan huruf 'E'.

Sifat-sifat Di peringkat Pertengahan - Sifat di peringkat ini diambil dan ditakrifkan dari ciri *sudut* (lihat gambarajah 3) yang dimiliki oleh huruf-huruf tertentu. Walau pun sudut ini adalah merupakan ciri yang asas seperti sifat-sifat yang ditunjuki di peringkat rendah di atas, tetapi oleh kerana bentuk daripada sudut ini berubah dari satu huruf ke huruf yang lain dan penentuannya agak kompleks maka ia lebih sesuai dikelaskan di dalam peringkat pertengahan. Sekali lagi huruf 'E' contohnya, ia mempunyai 2 - sudut.



Gambarajah 3: Sifat-sifat Diperingkat Rendah dan Pertengahan

Sifat-sifat Di peringkat Tinggi - Sifat-sifat yang dimiliki oleh huruf-huruf selain daripada bentuk-bentuk yang dibincangkan di atas adalah lebih kompleks. Ianya ditentukan daripada ciri-ciri *garis-lengkung* dan *orientasi garis*. Misalnya huruf 'O' terbentuk dari ciri garis-lengkung yang 'sempurna'. Setengah-setengah huruf itu pula terbentuk dari separuh garis-

lengkung/bulatan yang dibatasi oleh sama ada titik-akhir, titik-T atau sudut. Garis yang membentuk huruf dilihat dari segi orientasinya: garis-lurus atau garis-condong. Lihat di dalam gambarajah 4 untuk penerangan di atas.

Orientasi Garis Garis-lengkung

Gambarajah 4: Sifat-sifat Diperingkat Tinggi

4.4 Fasa Klasifikasi Kata-putus

Di dalam fasa yang akhir ini, satu sistem yang mempunyai kepintaran yang mencukupi bagi membuat keputusan terhadap calun imej hendaklah dibangunkan. Cadangan yang diutarakan di sini ialah dengan menyediakan satu 'kamus' yang menyimpan keseluruhan sifat-sifat yang dimiliki oleh setiap huruf-huruf. Sifat-sifat ini: titik-akhir, titik-T, sudut, orientasi-garis dan garis-lengkung, mestilah unik bagi setiap huruf-huruf.

Proses ini memerlukan suatu struktur data yang sesuai lagi efektif untuk digunakan di dalam 2 langkah berikut;

a. Merentasi calun imej bagi tujuan mengumpul kesemua sifat-sifat yang mungkin yang dimiliki olehnya.

b. Membangunkan kamus bagi sifat-sifat huruf dari A hingga Z.

Jika calun imej mempunyai sifat-sifat yang sama dengan sifat-sifat yang dimiliki oleh huruf sasaran di dalam kamus, maka proses pengecaman ini dikatakan berjaya. Jika proses pemadanan ini gagal, iaitu sifat-sifat ini tidak dijumpai di dalam kamus, maka calun ini diabaikan dan mesej kegagalan dalam mengecami calun itu haruslah diberikan.

5. Ciri-ciri Stesyen-kerja Pemerosesan Imej

Stesyen-kerja yang umum di dalam pembangunan sesuatu sistem pengecaman teks ini biasanya terdiri dari satu komputer induk (kemungkinan satu PC) yang dibekalkan dengan ciri-ciri berikut: 1 - Kerangka Penimbal, 1 - Monitor Pamiran, 1 - Pendigit Video, 1 - Punca Video.

Kerangka penimbal adalah terdiri dari ingatan video berkelajuan tinggi dan dibekalkan dengan sistem pamiran video. Kegunaannya ialah sebagai tempat simpanan data imej sebelum dan selepas pemerosesan di dalam ingatan utama komputer.
Seperti yang telah dibincangkan di dalam tajuk fasa pengambilan data di atas, input bagi video di dalam kes ini adalah diambil dari alat pengimbas.

Pendekatan yang mudah bagi penyelesaian terhadap pemerosesan imej adalah dengan membiarkan CPU melakukan semua tugas-tugas yang membabitkan kerja-kerja pembangunan imej. Sistem seperti ini adalah baik digunakan di dalam sistem berpesekitaran-tunggal, sebagai contoh sistem yang menggunakan pesekitaran DOS. Kemampuan bagi contoh sistem seperti ini bergantung semata-mata kepada kemampuan CPU itu sendiri, jalur-lebar busnya, dan antara-muka bagi kerangka penimbal.

Walaupun pendekatan ini memerlukan kemampuan minima yang boleh diterima pada kos yang paling rendah akan tetapi stesyen-kerja yang berdasarkan kepada mesin 80386 dan RISC(Reduced Instruction Set Computer) dan sistem bus yang boleh beroperasi pada kadar 33MHz adalah lebih sesuai.

6. Kesimpulan

Di dalam kertas-kerja ini, kita telah membincangkan secara sepintas lalu tentang bidang pengecaman teks dan kegunaannya. Seterusnya, dibincangkan tentang ciri-ciri penting yang harus ada baik dari segi keperluan terhadap perisian maupun terhadap perkakasan di dalam membangunkan sesuatu sistem pengecaman teks tulisan tangan. Masalah utama yang timbul di dalam proses pengecaman ini adalah datang dari ketidak-seimbangan terhadap penulisan. Bentuk huruf yang ditulis tangan adalah berbedza dari seorang penulis ke penulis yang lain. Walaupun alat pengecam optik yang terbaik iaitu mata, kadang-kadang tidak dapat mengecami setengah-setengah tulisan tangan, maka kerumitan ini sudah pasti lebih lagi dirasai oleh sistem pengecaman yang hendakdibangunkan. Masalah ini dapat dikurangkan jika teknik pelicinan dan pengecilan imej di dalam proses pra-pemerosesan dapat direka dengan lebih baik dan berkesan lagi sehingga ia dapat menerbitkan sifat-sifat imej yang lebih ketara lagi.

Rujukan

David D. Kerrick and Alan C. Bovik, 1988. *Microprocessor-based Recognition Of Handprinted Characters From A Tablet Input, Pattern Recognition, Vol 21, No. 5 pp. 525 - 537.*

David Essex, August 89. *Scanners Enter New Age, PC Resource.*

Gene Smarte, Walt Penny, Bobby Saffari, December 1989. *...Sound and Image Processing, Byte .*

J. Mantas, 1986. *An Overview Of Character Recognition Methodologies, Pattern Recognition, Vol 19, No. 6,pp 425-430.*

K. S. Fu, 1976. *Digital Pattern Recognition, Springer-Verlag .*

Pierre A. Devijver, Josef Kittler, 1987. *Pattern Recognition Theory and Applications, NATO ASI Serries.*

R. M. Brown, T. H. Fay and C. L., 1988. *Walker, Handprinted Symbol Recognition System, Pattern Recognition, Vol. 21, No. 2, pp. 91 - 118.*

Raymond W. Smith, 1987. *Computer Processing Of Line Images: A Survey, Pattern Recognition, Vol. 20, No. 1,pp 7-15*

Simon Kahan, Theo Pavlidis, Hendry S. Baird, March 1987. *On the Recognition of Printed Characters of Any Font and Size, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. PAMI-9, No. 2.*

Sing Tze Bow, 1984. *Pattern Recognition, Dekker .*

Yat Keung Chu, 1986. *An Alternative Smothing And Stripping Algorithm for Thinning Digital binary Patterns, signal Processing 11 (1986) 207 - 222 .*