BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pengolahan citra digital adalah manipulasi dan interprestasi digital dari citra dengan bantuan komputer. Salah satu pengolahan citra yaitu pengenalan huruf atau bisa disebut OCR (*Optical character Recognition*). OCR (Optical Character Recognition) adalah aplikasi yang berfungsi untuk men-*scan* gambar dan dijadikan teks. Dengan adanya OCR, *image* yang bertulisan tangan, tulisan mesin ketik atau *computer text*, dapat dimanipulasi. Teks yang di-*scan* dengan OCR dapat dicari kata per-kata atau per-kalimat. Dan setiap teks dapat dimanipulasi, diganti, atau diberikan barcode[1]. OCR diperlukan untuk dokumen karya tulis ilmiah yang tidak memiliki salinan *softcopy*, hal tersebut bisa dikarenakan kehilangan, kerusakan file, ataupun dokumen tersebut memang hanya memiliki salinan *hardcopy* saja. Selain itu, satu-satunya pilihan untuk mendigitalkan dokumen kertas tercetak adalah dengan mengetik ulang teks secara manual, tetapi hal ini bisa menghabiskan banyak waktu serta bisa menyebabkan ketidakakuratan dan kesalahan pengetikan.

Beberapa metode yang pernah digunakan untuk OCR yaitu dengan menggunakan *Backpropagation*, Jaringan saraf tiruan[4], dan SVM[26]. Penelitian lainnya tentang OCR dilakukan oleh Fajry Hamzah, yaitu mengenai pengenalan tulisan dan ekstraksi informasipada citra abstrak skripsi menggunakan *support vector machine* dan *rules based system* menghasilkan tingkat akurasi yang kecil yaitu sebesar 5,02% untuk *case sensitive* dan 5,47% untuk *case insensitive*, sedangkan untuk tingkat akurasi pengenalan karakter menggunakan SVM itu sendiri mencapai 46.61% untuk *case sensitive* dan 54.30% untuk *case insensitive*. Rendahnya tingkat akurasi pengenalan pada citra abstrak dipengaruhi oleh proses segmentasi yang kurang mampu menyelesaikan masalah yang ada pada pemisahan karakter dan kurangnya metode dalam mengekstraksi ciri citra. Sementara itu, ekstraksi fitur zoning dan SVM dengan menggunakan kernel linear

telah digunakan untuk mengekstraksi ciri citra aksara sunda dengan menghasilkan tingkat akurasi sebesar 99,75%[24].

SSVM merupakan pengembangan *smoothing technique* yang menggantikan *plus function* SVM dengan integral dari fungsi signoid *neural network*. SSVM memiliki performa yang lebih baik dalam mengatasi data berdimensi tinggi dan data jumlah besar juga memiliki *running* yang lebih cepat dan akurasi yang lebih besar[6].

Dari penjelasan di atas, berdasarkan beberapa penelitian tersebut, dapat disimpulkan bahwa metode *Smooth Support Vector Machine* (SSVM) dapat digunakan untuk pengklasifikasian citra, namun bergantung kepada metode *preprocessing* dan ekstraksi fitur yang digunakan. Oleh karena itu, pada penelitian ini akan menggunakan metode *Smooth Support Vector Machine* (SSVM) untuk melakukan *Optical Character Recognition* (OCR) pada dokumen karya tulis ilmiah dengan menggunakan ekstraksi fitur zoning, yang selanjutnya diharapkan sistem dapat mengenali karakter dengan baik dan tingkat akurasi yang dihasilkan tinggi.

1.2 Identifikasi Masalah

Adapun identifikasi masalah pada penelitian ini adalah rendahnya tingkat akurasi yang diperoleh dari penelitian sebelumnya tentang OCR pada citra dokumen abstrak skripsi.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah membangun sistem OCR dengan SSVM. Adapun tujuan dari penelitian ini yaitu mengukur performa dari implementasi SSVM dan menghitung akurasi yang diperoleh pada kasus pengenalan karakter dokumen karya tulis ilmiah.

1.4 Batasan Masalah

Agar penelitian yang dilakukan lebih terarah sesuai dengan tujuan yang diinginkan, makan berikut adalah batasan masalah yang digunakan.

1. Data Masukan

   Data masukan yang digunakan yang digunakan pada penelitian ini dibagi menjadi 2 bagian yaitu data latih dan data uji.

   a. Data latih

1) Citra dokumen hasil *scan* atau foto berformat (.png).
2) Data latih diambil dari 80 citra karakter alfabet yaitu huruf kapital (A-Z), huruf kecil (a-z), angka (0-9) dan beberapa simbol lainnya berupa .,:;/[]%*'()?!#+-=
3) Ukuran citra yang digunakan yaitu 16x16 piksel.
4) Gaya *font* yang digunakan hanya berjenis *Times New Roman* dengan *font size* 16.

b. Data uji

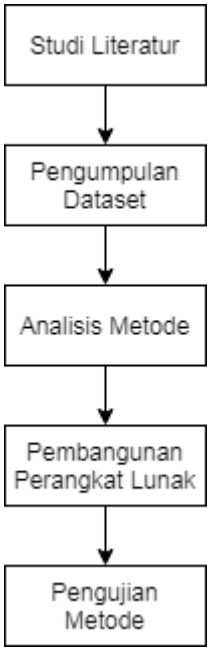
Data uji yang digunakan yaitu gambar hasil *scan* dokumen karya tulis ilmiah skripsi yang mengandung karakter (kecuali bagian yang mengandung rumus, gambar, dan tabel) dalam format berupa .jpg/ .png/ .jpeg.

2. Proses
   a. Metode *preprocessing* yang digunakan yaitu *grayscale*, binerisasi, *skew corrrection*, segmentasi baris dan karakter, *resize*, serta ekstraksi fitur.
   b. Ekstraksi fitur yang digunakan yaitu *Zoning Image Centroid Zone (ICZ)* untuk memisahkan ciri-ciri yang terdapat pada setiap karakter.
3. Data Keluaran

Keluaran yang dihasilkan yaitu berpua teks digital yang akan disimpan dalam file berformat (.txt).

1.5 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah metode kuantitatif [23]. Metode ini digunakan karena data yang digunakan dalam penelitian ini berbentuk angka dan bersifat fakta serta bisa diukur secara akurat dengan alat yang objektif sehingga bisa membantu dalam penelitian ini.



Gambar 1.1 Alur Penelitian

Adapun penjelasan langkah-langkah yang dilakukan dalam penelitian ini adalah sebagai berikut.

1.5.1 Studi Literatur

Studi ini dilakukan dengan cara mempelajari, meneliti dan menelaah berbagai literatur-literatur yang bersumber dari buku-buku, teks, jurnal ilmiah, situs-situs di internet, dan bacaan-bacaan yang terkait dengan topik pengenalan karakter, metode dalam pengolahan citra, metode ekstraksi fitur, dan metode klasifikasi SSVM.

1.5.2 Pengumpulan Dataset

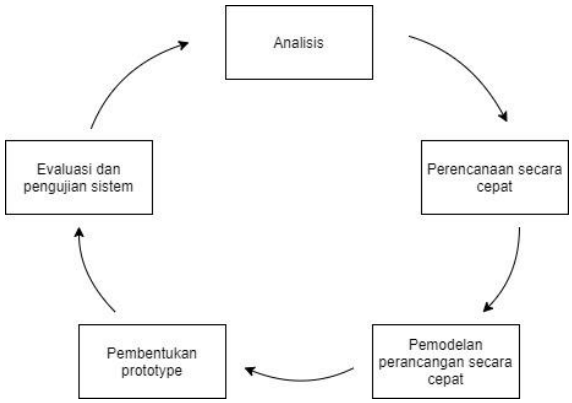
Dataset yang akan dikumpulkan disini terbagi menjadi dua, yaitu dataset untuk data latih dan dataset untuk data uji. Data latih yaitu berisi data citra karakter yang berjumlah sebanyak 80 gambar yang terdiri dari huruf A-Z, a-z, angka 0-9, serta karakter tambahan berupa simbol .,:;/[]%*'()?!#+-=. Sementara data uji adalah data hasil *scan* dokumen karya tulis ilmiah skripsi mahasiswa Teknik Informatika Universitas Komputer Indonesia. Dari data tersebut akan dikumpulkan sebanyak 40 hasil *scan* dokumen yang akan digunakan sebagai data uji.

1.5.3 Analisis Metode

Selanjutnya setelah dataset terkumpul kemudian masuk ke dalam tahap analisa,dimana tahap ini data literatur dan referensi yang sudah terkumpul dipelajari terkait metode yang akan digunakan yaitu SSVM, dan fitur apa saja yang dibutuhkan dan hal lainnya terkait pembangunan sistem termasuk menganalisa bagaimana sistem akan berjalan mulai dari tahap *input* kenudian proses hingga hasil berupa teks hasil pengenalan sistem.

1.5.4 Pembangunan Perangkat Lunak

Dalam penelitian ini pembangunan perangkat lunak yang digunakan adalah model prototipe, karena metode ini menyajikan gambaran lengkap dari suatu sistem perangkat lunak yang akan dibangunnya[7]. Berikut ini adalah proses dari model prototipe pada gambar 1.2.



Gambar 1.2 Model Prototipe

Berikut ini adalah langkah-langkah yang dilakukan dalam proses Model Prototipe.

1. Analisis

   Tahap - tahap ini merupakan kegiatan untuk menganalisis dan mengumpulkan kebutuhan data dengan cara membaca buku dan jurnal tentang metode yang digunakan dalam preprocessing citra tulisan tangan. Selain itu mengumpulkan sampel tulisan tangan.

2. Perencanaan Secara Cepat

   Tahap - tahap ini merupakan perancangan awal dari aplikasi yang dibangun, seperti perancangan antarmuka aplikasi. Perancangan dibuat

berdasarkan dengan data-data yang telah dikumpulkan pada tahap pengumpulan kebutuhan dan analisis.

3. Pemodelan perancangan secara cepat

   Tahap - tahap ini merupakan pembuatan tampilan dan bentuk program.

4. Pembentukan Prototype

   Mengimplementasikan perencanaan dan pemodelan menjadi bentuk prototipe perangkat lunak untuk sistem OCR mulai dari tahap *preprocessing, training dan testing.*

5. Evaluasi dan pengujian Sistem

   Perangkat lunak yang telah dibangun akan dievaluasi jika terdapat kekurangan atau kesalahan kode program. Selain itu, akan dilakukan pengujian terhadap fungsionalitas sistem OCR dengan menggunakan *Black-Box*.

1.5.5 Pengujian Metode

Perangkat lunak yang telah dibuat dan diuji secara menyeluruh akan masuk ke tahap pengujian metode agar dapat mengetahui tingkat akurasi yang dihasilkan SVM untuk kasus pengenalan karakter pada dokumen karya tulis ilmiah. Hasil yang telah dikeluarkan oleh perangkat lunak akan dihitung akurasi ketepatan pengenalannya menggunakan metode *Classification Accuracy* sehingga bisa diketahui rata-rata tingkat akurasi pengenalan karakter yang dihasilkan.

1.6 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran secara umum mengenai permasalahan dan pemecahannya. Sistematika penulisan skripsi ini adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini berisi uraian tentang latar belakang masalah, identifikasi masalah, maksud dan tujuan, batasan masalah pada penelitian ini.

BAB 2 LANDASAN TEORI

Pada bab ini diuraikan secara teoritis hal – hal yang digunakan dalam penelitian, seperti pengolahan citra digital, metode – metode yang digunakan pada

tahap *preprocessing*, pengenalan *Smooth Support Vector Machine* (SSVM) sebagai metode klasifikasi, bahasa pemrograman yang digunakan, dan metode pengujian,

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi tentang analisis dan perancangan sistem yaitu meliputi analisis masalah, analisis data masukan, analisis proses *preprocessing*, ekstraksi fitur dan klasifikasi, analisis data keluaran, analisis kebutuhan fungsional dan *non* fungsional, perancangan antar muka, struktur menu dan pesan, serta jaringan semantik.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini berisi tentang hasil dari keseluruhan tahap analisis dan perancangan yang meliputi implementasi data masukan, implementasi perangkat keras dan perangkat lunak yang digunakan, implementasi antarmuka, serta pengujian dan hasil pengujian fungsionalitas sistem dan pengujian akurasi metode.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini akan diuraikan hasil dari penelitian yang telah dilakukan sesuai dengan tujuan yang sudah ditetapkan, disertai dengan saran untuk peneliti selanjutnya agar penelitian selanjutnya menjadi lebih baik lagi.