## ABSTRAK

Pada suatu organisasi maupun instansi penggunaan media surat dinilai lebih informatif dalam menyampaikan informasi yang bersifat resmi dan penting. Surat resmi dari suatu instansi memiliki nilai hukum dan dapat dijadikan sebagai alat bukti historis, sehingga pengagendaan surat-surat tersebut perlu dilakukan. Teknologi OCR (*optical character recognition*) dengan *Tesseract OCR Engine* dapat diterapkan untuk membantu proses pendataan agenda surat tersebut, yaitu dengan melakukan proses ekstraksi informasi dari citra dokumen surat yang kemudian akan menghasilkan data digital untuk dapat diolah sesuai kebutuhan. Namun hasil akurasi *Tesseract* akan berkurang apabila terdapat objek gangguan pada citra yang akan diproses.

Solusi dari permasalahan ini dapat diatasi dengan menambahkan *image preprocessing* sebelum proses pengenalan oleh *Tesseract* untuk meningkatkan kualitas citra, sehingga hasil pengenalan karakter menjadi lebih baik. Tahapan *image preprocessing* yang dilakukan yaitu *scaling*, *brightness*, g*rayscale*, *gaussian filtering*, *otsu thresholding* dan erosi. Selanjutnya data hasil pengenalan OCR akan diklasifikasikan menggunakan algoritma *Regular Expression.* Adapun data yang diambil adalah data nama instansi pengirim, nomor surat, tanggal surat, perihal dan penerima surat.

Hasil yang didapat dari pengujian yang telah dilakukan pada 15 dokumen surat dengan format penulisan yang berbeda, akurasi *Tesseract* OCR dapat ditingkatkan dengan menerapkan *image preprocessing*. Nilai rata-rata akurasi untuk pengenalan karakter pada citra surat tanpa *image preprocessing* sebesar 73,503388%, sedangkan pada citra surat dengan *image preprocessing* menghasilkan nilai akurasi rata-rata sebesar 90,58362%. Sehingga nilai peningkatan rata-rata yang dihasilkan pada pengenalan objek citra surat yaitu sebesar 17,08023%. Selain itu algoritma *Regular Expression* dapat digunakan untuk proses klasifikasi data dengan nilai rata-rata akurasi yang dihasilkan untuk atribut nama instansi pengirim sebesar 91,26785%, atribut nomor surat sebesar 90,36643%, atribut tanggal surat sebesar 93,25311%, atribut perihal sebesar 91,69876% dan untuk atribut penerima surat sebesar 92,89901%.



**Kata Kunci** : Surat, *Optical Character Recognition, Tessercat OCR Engine, Regular Expression*