ISSN: 2477-5126
e-ISSN: 2548-9356

Implementasi *Optical Character Recognition* berbasis *Deep Learning* untuk Ekstraksi Data Sertifikat Tanah

Dinar Nugroho Pratomo [1*)], Diyah Utami Kusumaning Putri[2], Azhari[3]
[1]Teknologi Rekayasa Perangkat Lunak, Sekolah Vokasi, Universitas Gadjah Mada
[2,3]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada
[1]Sekip Unit 1, Caturtunggal, Depok, Sleman, Daerah Istimewa Yogyakarta, 55281
[2,3]Sekip Utara Bulaksumur 21, Sinduadi, Mlati, Sleman, Daerah Istimewa Yogyakarta, 55281
email: [1]dinar.nugroho.p@ugm.ac.id, [2]diyah.utami.k@ugm.ac.id, [3]arisn@ugm.ac.id

Abstract *- Data owned by BPN in the form of copies of land certificates have been collected in the safe deposit box at each regional BPN for many years. There are many cases where the taxes that should be borne by landowners may not be paid due to lack of data updates. This problem is that there is no proper digitization of data. The research activity begins with collecting samples of several copies of land certificates from several BPNs in the Brebes area, then pre-processing the data (pre-processing) with 3 stages, namely scaling, greyscalling, and binarization. The method used is OCR and CNN to extract the data obtained. Tests are performed on every important page of the certificate. These pages are the cover section, first registration, letter of measurement, quotation and list of books. c. each page has a different data extraction result according to the model labeling that has been done. The average percentage of conformity results is 97.05%. This proves that the method used is reliable.*

Abstrak **–** Data yang dimiliki BPN berupa salinan sertifikat tanah sudah terhimpun di brankas penyimpanan pada setiap BPN daerah bertahun-tahun lamanya. Banyak kasus yang dapat membuat pajak yang seharusnya ditanggung oleh pemilik tanah bisa jadi tidak terbayarkan karena kurangnya pembaharuan data. Permasalahan ini adalah belum adanya pendigitalisasian data dengan baik. Kegiatan penelitian diawali dengan pengumpulan contoh beberapa salinan sertifikat tanah dari beberapa BPN daerah Brebes, kemudian dilakukan pra-pemrosesan data (pre-processing) dengan 3 tahapan yaitu scalling, greyscalling, dan binarization. Metode yang digunakan menggunakan OCR dan CNN untuk mengekstraksi data yang diperoleh. Pengujian dilakukan di setiap halaman penting pada sertifikat. Halaman tersebut adalah bagian cover, pendaftaran pertama, surat ukur, kutipan dan daftar buku c. setiap halaman memiliki hasil ekstraksi data yang berbeda sesuai dengan pelabelan model yang sudah dilakukan. Rata-rata prosentase hasil kesesuaian sebesar 97,05%. Hal ini membuktikan bahwa metode yang digunakan sudah sangat handal.

Kata Kunci **– Sertifikat tanah, BPN, CNN, OCR, Pre-processing**

*) **penulis korespondensi**: Dinar Nugroho Pratomo
Email: dinar.nugroho.p@ugm.ac.id

I. PENDAHULUAN

Pendataan sertifikat tanah saat ini belum terdigitalisasi dengan baik oleh Badan Pertanahan Nasional (BPN). Data yang sudah terhimpun berupa salinan sertifikat tanah yang disimpan di brankas penyimpanan BPN setiap daerah bertahun-tahun lamanya. Hal ini mengakibatkan data tidak terupdate dengan baik. Misalnya ketika pemilik tanah meninggal dan harus digantikan dengan ahli waris, apabila ahli waris tersebut tidak mengurus balik nama maka sertifikat tersebut tidak akan terupdate. Penyimpanan secara manual ini juga dapat mengakibatkan file rusak.

Permasalahan ini dapat menjadi latar belakang permasalahan lain di bidang perpajakan. Pajak yang seharusnya ditanggung oleh pemilik tanah bisa jadi tidak terbayarkan karena pemilik tanah sudah meninggal dan ahli waris tidak mengurusnya dalam pergantian sertifikat.

Permasalahan yang dimiliki oleh BPN adalah belum adanya pendigitalisasian data dengan baik. Upaya yang dapat dilakukan yaitu membuat suatu sistem yang dapat mengekstraksi informasi dari sertifikat tanah yang di-scan sehingga petugas tidak perlu memasukkan data satu per satu ke dalam database. Sistem ini diharapkan mampu membantu BPN untuk melakukan digitalisasi sertifikat tanah sehingga dapat membentuk suatu database yang dapat digunakan untuk kepentingan lain di masa yang akan datang.

Ekstraksi data penting yang terdapat pada sertifikat tanah di setiap halamannya berdasarkan tata letak atau template yang sudah dimodel berdasarkan label yang sudah ditentukan dan menyesuaikan isi keterangan terkait menggunakan Optical Character Recognition (OCR). OCR adalah bidang yang aktif dan terus berkembang selama beberapa dekade terakhir[1]. OCR merupakan sebuah aplikasi yang dapat menerjemahkan *image character* kedalam bentuk teks melalui penyesuaian pola-pola dari karakter setiap barisnya terhadap pola yang terdapat pada penyimpanan di sistem[2]. OCR dapat menyelesaikan masalah seperti mengekstraksi informasi atau pembacaan tulisan dalam sebuah *image*[3]. Bagian yang harus diperhatikan dalam pengembangan OCR adalah mengekstrak ciri dan mengenali pola[4]. Ekstraksi ciri bertujuan agar didapatkan ciri atau identitas dari sebuah karakter[5][6]. Pengenalan pola berfungsi sebagai pencocokan pola berdasarkan input dengan pola yang ada di *knowledge base*[7]. Pendekatan *deep learning* dapat digunakan untuk mengenali pola dan klasifikasi dalam model yang dilatih. CNN digunakan dalam visi komputer untuk klasifikasi gambar dan pengenalan objek[8]. CNN menggambarkan variasi dari *multilayer perceptron* yang dapat beroperasi pada data dua

dimensi dengan cara kerja mirip dengan jaringan syaraf yang ada pada manusia[9]. CNN menunjukkan kinerja yang menonjol dalam penglasifikasian citra, anotasi gambar, dan berbagai bidang visi komputer lainya[10][11]. CNN mampu mengenal karakter baru yang sebelumnya tidak ada dalam dataset dan dapat melakukannya dengan lebih ringkas[11]. Hal ini disebabkan pada CNN terdapat *feature extraction layer*.

Metode CNN cocok dikombinasikan dengan OCR adar mendapatkan hasil yang lebih optimal. Algoritma CNN dapat mengenali karakter baru yang sebelumnya tidak ada dalam dataset dan dapat melakukannya dengan lebih ringkas karena di dalam CNN itu sendiri sudah terkandung Feature Extraction Layer[12]. Penelitian ini, juga menggunakan library pada python yaitu keras, agar CNN dapat dijalankan dengan tepat, efektif dan lebih ringkas. Secara umum, dari penelitian ini, diharapkan dapat memberikan pengetahuan baru mengenai penerapan OCR dan CNN dalam mengekstraksi data yang ada pada lembar sertifikat tanah ke dalam bentuk digital.

II. PENELITIAN YANG TERKAIT

Algoritma yang efektif untuk mengesktrak karakter dari gambar (dokumen yang di-scan) yang dibuat dengan mengambil gambar berwarna sebagai input[13]. Tujuan utama pada Penelitian terdahulu yang dilakukan oleh Robby et al. adalah untuk secara otomatis mengekstraksi teks dan menampilkan informasinya. Jadi, sangat berguna untuk mendeteksi, mengekstrak dan mengenali teks dari gambar digital dan mengotomatiskan ekstraksi dokumen teks, atau metadata dari gambar digital, untuk memanfaatkan pengambilan database gambar dengan benar. Metode utama yang digunakan pada sistem yaitu konversi *grayscale*, deteksi *Canny-edge, adaptive thresholding, morphological operation*, deteksi blok karakter (juga kata dan baris), *optical character recognition*[13]. Metode yang diusulkan pada penelitian tersebut ternyata tidak cocok untuk semua jenis gambar digital. Metode tersebut tidak mampu menangani gambar yang banyak memiliki karakter atau simbol yang tidak dikenali. Sistem ini hanya mampu menangani ekstraksi gambar yang sebagian besar berisi teks sebagai dokumen.

Penelitian lain yang dibuat oleh Sukanya et al. membuat suatu *tool* untuk mengekstraksi teks dari dokumen gambar yang di-scan dan mengkonversinya ke format yang dapat diedit. Gambar disegmentasi menjadi karakter dengan menggunakan *connected components* (CC) dan rekombinasi tepi menggunakan stroke width[14]. Gambar ini kemudian dikonversi ke format yang dapat diedit dengan menggunakan teknologi OCR dan *maximally stable extremal region* (MSER) untuk segmentasi. Sistem yang diusulkan juga dapat mengekstraksi objek dari gambar dengan menggunakan Jaringan Syaraf Tiruan (JST). Tool ini dikembangkan menggunakan MATLAB dan hasilnya disimpan dalam variabel atau dapat pula diekstrak ke dokumen yang dapat diedit. Kinerja sistem diukur dengan menggunakan dua parameter yaitu tingkat precision dan tingkat *recall* dan memiliki sekitar 88% *precision* dan tingkat *recall* 97% yang lebih tinggi dari sebagian besar metode yang diusulkan sebelumnya.

Jha et al. mengimplementasikan sistem otomatis yang mengekstrak rincian yang relevan pada formulir bank seperti nama penerima pembayaran, jumlah, tanggal, nama bank menggunakan OCR dan *deep learning* dan memverifikasi tanda tangan pada cek dengan tanda tangan yang disimpan dalam database menggunakan ekstraksi fitur dan principal component analysis[15]. Sistem yang diusulkan menggunakan convolution neural network yang dimodifikasi untuk mengekstraksi konten tulisan tangan pada formulir dimana dalam dataset IAM digunakan untuk melatih model dan mendapatkan hasil optimal. Sistem ini akan memfasilitasi proses dan mengarah pada pengurangan waktu dan biaya. Efisiensi dan kinerja diukur pada sekumpulan data formulir bank yang dihasilkan sendiri.

Penelitian Ishchenko et al. mengusulkan metode ekstraksi area teks pada gambar dokumen yang di-scan menggunakan *linear filtering* dan *threshold image transformation*[16]. *Linear filtering* memungkinkan Anda untuk menghaluskan (*smoothing*) nilai intensitas piksel di dalam area yang homogen[17]. Threshold digunakan untuk mengisolasi area yang homogen pada gambar yang membentuk fragmen teks dari background. Pengujian metode dilakukan untuk segmentasi gambar tekstual dari arsip surat kabar yang di-scan dari database dokumen MediaTeam di Universitas Oulu (Finlandia). Metode yang diusulkan dapat meningkatkan kualitas pemilihan area ini dibandingkan dengan metode sebelumnya.

III. METODE PENELITIAN

Pada tahap awal penelitian dilakukan pengumpulan dan pengkajian referensi dari buku, jurnal, karya ilmiah artikel terkait macam sertifikat tanah, Optical Character Recognition (OCR), teori neural network, teori deep learning, dan teori Convolutional Neural Network (CNN). Pengumpulan data akan dilakuukan dengan meminta data dari Badan Pertanahan National (BPN) kabupaten Brebes berupa fotocopy sertifikat dan hasil digital dari foto dan scan. Sumber data lain juga didapatkan dari kantor BPN tegal dan Lembaga daerah yang memiliki contoh jenis sertifikat tanah yang lain.

A. Praproses data

Data hasil yang dikumpulkan memiliki cacat, kabur, atau warna yang terlalu kontras dan sebagainya sehingga informasi yang terkandung dalam gambar tidak dapat ditafsirkan dengan benar. Jadi tugas pertama diperlukan untuk mendenoise gambar. Ini disebut pra-pemrosesan [11]. Pra-pemrosesan gambar adalah tahapan dalam pengenalan pola untuk meningkatkan kualitas gambar yang diperoleh. Gambar digital pertama kali akan masuk ke dalam tahap preprocessing. Tahap ini dibagi kedalam 3 tahapan kecil lainnya, 3 tahapan tersebut adalah scaling, greyscaling, dan binarization.

Pada tahap scaling semua gambar yang memiliki ukuran panjang lebih dari 1024 akan dikecilkan ke panjang 1024 dengan ratio ukuran yang sama. Sebagai contoh masuk gambar 3264x2448 (ratio 43). gambar tersebut akan diubah menjadi ukuran 1024x768 (ratio 4:3). Scaling dilakukan untuk mempercepat proses yang dilakukan, karena semakin besar gambar maka akan semakin banyak pixel yang perlu

diproses. Gambar yang telah melewati tahap scaling kemudian akan dikonversi menjadi gambar abu- abu (greyscaling). Secara sederhana tahap ini akan mengubah gambar berwarna yang memiliki 3 chamel warna menjadi gambar abu-abu yang memiliki I channel warna.

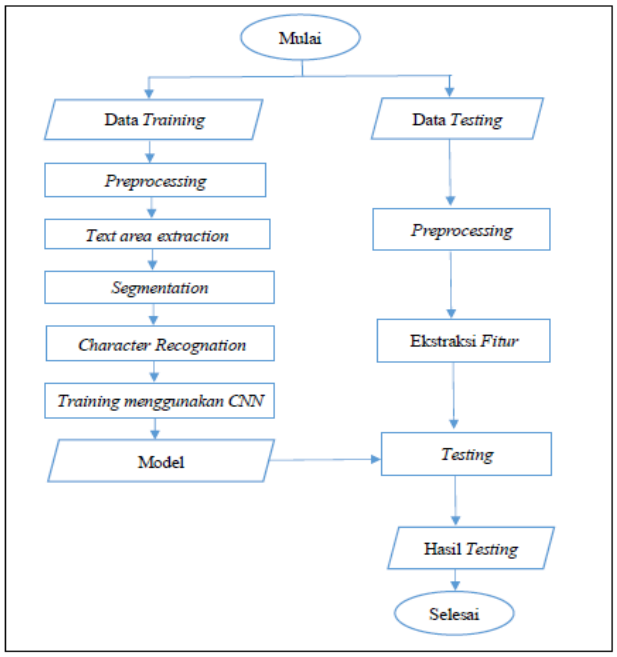
Tahap greyscaling dilakukan untuk mendapatkan hasil yang lebih optimal saat tahap binarization dilakukan (binarization hanya akan membaca 1 channel warna). Terdapat 8 algoritma umum yang digunakan pada penelitian ini. yaitu: *average* (a), *luminance* (b), *desaturation* (c), *decomposition maximum* (d) *decomposition minimum* (e), *single chanmel color red* (f). *single channel color green* (g), *single channel color blue* (h).

B. Perancangan sistem

Pada tahap ini akan dilakukan perancangan pembuatan sistem. Sistem diawali dengan pembagian dataset yang telah dikumpulkan menjadi dataset training dan dataset testing. Kemudian dilanjutkan dengan pra-pemrosesan (preprocessing) data dengan melakukan grayscaling, rotation, menghilangkan noise (filtering) dan menebalkan tulisan (dilation/ skeletonizing), cropping, resizing, dan thresholding. Setelah itu, proses dilanjutkan dengan ekstraksi fitur mengunakan OCR.

Gambar 1. Pelabelan data untuk menentukan lokasi text

Hasil ekstraksi selanjutnya digunakan sebagai modeling label untuk menentukan lokasi data yang akan diambil beserta menampilkan isi label sebagai data training seperti yang ditunjukan pada Gambar 1. Penelitian ini menggunakan metode (Convolutional Neural Network (CNN) untuk proses pelatihan (training). Model yang didapatkan dari proses pelatihan akan diuji menggunakan data testing untuk mengevaluasi kinerja sistem dengan beberapa metode pengukuran. Perancangan sistem dijelaskan dengan diagram alir pada Gambar 2.



Gambar 2. Diagram Alir Penelitian

IV. HASIL DAN PEMBAHASAN

Uji coba pada Penelitian ini dilakukan pada sertifikat tanah yang berupa sertifikat hak milik, sertifikat hak guna bangunan dan petok. Halaman pada sertifikat tersebut yang diolah adalah bagian cover, pendaftaran pertama, surat ukur, kutipan dan daftar buku c. setiap halaman memiliki hasil ekstraksi data yang berbeda sesuai dengan pelabelan model yang sudah dilakukan.

A. Hasil Akurasi

Salah satu contoh hasil akurasi sertifikat hak milik yang ditunjukan pada tabel 1. Akurasi tersebut didapat dengan mengacak data testing dan dilakukan pengulangan percobaan sebanyak 30 kali.

Tabel I. Prosentase hasil ekstraksi data pada halaman cover

Label	Total karakter pada citra	Prosentase kesesuaian
Kode	270	100%
No	150	100%
kecamatan	330	100%
desa	330	100%
Isian 307	450	98,82%
Isian 208	450	99,73%
no seri	57	90,17%
Rata-rata keberhasilan		98,39%

Table 1 menunjukan hasil hasil yang tinggi karena ekstraksi data yang didapat adalah text dari mesin ketik sehingga, akurasi yang besar didapat karena karakter yang dideteksi mudah untuk dibaca. Masalah hanya pada nomor seri, karena pengelompokan yang dilakukan berdasarkan isian nomor pada kotak-kotak yang ada, dan beberapa kesalahan terjadi karena pengenalan karakter tanda titik ".".

Secara keseluruhan hasil prosentase sesuai dengan halaman yang diuji ditampilkan pada tabel 2. Hasil yang ditampilkan menunjukan rata-rata keberhasilan dari prosentase kesesuaian

karakter yang dikenali untuk seluruh label yang ada pada setiap halaman yang dideteksi.

Tabel II. Prosentase hasil ekstraksi data pada bagian halaman

Label	Prosentase kesesuaian
cover	98,39%
pendaftaran pertama	95,13%
Surat ukur	98,82%
Pendaftaran peralihan	93,73%
daftar buku c	99,17%
Total rata-rata	97.05%

B. Hasil Akurasi

Hasil yang ditampilkan pada tabel 2 dapat diketahui bahwa rata-rata persentase keberhasilan pengenalan pada pengujian halaman cover 98,39%, pada halaman pendaftaran pertama sebesar 95,13%, pada halaman surat ukur sebesar 98,82%, pada halaman pendaftaran peralihan sebesar 93,73%, dan pada halaman daftar buku c sebesar 99,17%. Sedangkan total rata-rata persentase keberhasilan pengenalan secara menyeluruh dari seluruh halaman yang diuji sebesar 97,05%. Tingkat keberhasilan pengenalan yang dihasilkan cukup tinggi meskipun jenis dan ukuran huruf yang digunakan sebagai masukan berbeda dengan template.

V. KESIMPULAN

A. Kesimpulan

Penggunaan OCR dan CNN pada ekstraksi data untuk halaman-halaman yang ada pada sertifikat cukup handal. Hal ini terbukti dengan hasil prosentase kesesuaian sebesar 97,05%. Hal ini membuktikan bahwa hasil kesesuian yang didapat menggunakan metode OCR diberbasis metode deep learning yaitu CNN sangat handal terhadap perubahan model template yang sudah ditentukan. Dengan menggunakan data training yang baik dan optimal, maka subset dari data training tersebut juga akan menghasilkan hasil pendeteksian yang baik.

B. Saran

Penelitian ini masih dapat dikembangkan lebih lanjut. Penggunaan metode diatas belum diuji secara maksimal untuk sertifikat tanah dengan *hand writing.* Butuh tahapan lebih lanjut untuk mengekstraksi data tersebut. Selain itu data yang dituliskan dengan *hand writing* / tulisan tangan biasanya masih banyak terdapat singkatan yang kurang dimengerti. Selain itu *hand writing* biasanya tidak sesuai dengan template / kolom yang sudah disediakan.

DAFTAR PUSTAKA

[1] R. Anil, K. Manjusha, S. Kumar, Sachin, K. P. Soman, *"Convolutional Neural Networks for the Recognition of Malayalam Characters",* Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014", 2015.

[2] A. Setiawan, H. Sujaini, and A. Bijaksana Pn, "Implementasi Optical Character Recognition (OCR) pada Mesin Penerjemah Bahasa Indonesia ke Bahasa Inggris," J. Sist. dan Teknol. Inf., vol. 5, no. 2, pp. 135–141, 2017.

[3] H. Oktavianto and H. W. Sulistyo, "Optical Character Recognition Untuk Ekstraksi Teks Rambu Lalu Lintas," JUSTINDO (Jurnal Sist. Teknol. Inf. Indones., vol. 3, no. 1, pp. 15–21, 2018.

[4] S. S. Patil and A. S. Bhalchandra, "Pattern Recognition Using Genetic Algorithm," in 2017 International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 310–314, 2017.

[5] R. I. Borman and B. Priyopradono, "Implementasi Penerjemah Bahasa Isyarat Pada Bahasa Isyarat Indonesia (BISINDO) Dengan Metode Principal Component Analysis (PCA)," J. Pengemb. IT, vol. 03, no. 1, pp. 103–108, 2018.

[6] A. C. Roy, M. K. Hossen, and D. Nag, "License Plate Detection and Character Recognition System for Commercial Vehicles Based on Morphological Approach and Template Matching," Electr. Eng. Inf. Commun. Technol. (ICEEICT), 2016 3rd Int. Conf., pp. 1–6, 2016.

[7] S. Hartanto, A. Sugiharto, and S. N. Endah, "Optical Character Recognition Menggunakan Algoritma Template Matching Correlation," J. Masy. Inform., vol. 5, no. 9, pp. 1–11, 2015.

[8] A. Mulyanto, E. Susanti, F. Rossi, W. Wajiran, R. I. Borman, "Penerapan Convolutional Neural Network (CNN) pada Pengenalan Aksara Lampung Berbasis Optical Character Recognition (OCR)", Jurnal Edukasi dan Penelitian Informatika (JEPIN), Vol. 7, No. 1, pp. 52-57, 2021

[9] M. M. Taslim, K. Gunadi, and A. N. Tjondrowiguno, "Deteksi Rumus Matematika pada Halaman Dokumen Digital dengan Metode Convolutional Neural Network," J. Infra, vol. 7, no. 2, pp. 123–129, 2019.

[10] S. M. A. Sharif, N. Mohammed, N. Mansoor, and S. Momen, "A hybrid deep model with HOG features for Bangla handwritten numeral classification," in Proceedings of 9th International Conference on Electrical and Computer Engineering, ICECE 2016, 2017, no. February 2018, pp. 463–466.

[11] W. A. Saputra, M. Z. Naf'an, and A. Nurrochman, "Implementasi Keras Library dan Convolutional Neural Network Pada Konversi Formulir Pendaftaran Siswa," J. RESTI (Rekayasa Sist. Dan Teknol. Informasi), vol. 1, no. 10, pp. 524–531, 2019.

[12] A. Setiawan, H. Sujaini, and A. Bijaksana Pn, "Implementasi Optical Character Recognition (OCR) pada Mesin Penerjemah Bahasa Indonesia ke Bahasa Inggris," J. Sist. dan Teknol. Inf., vol. 5, no. 2, pp. 135–141, 2017.

[13] A. Robby, G. et al. "Extracting Text Information from Digital Images", International Journal of Scientific & Engineering Research, vol 10, issue 6, pp.1350-1356, 2019.

[14] S. Sukanya, S. J. Gladwin, C. V. Kumar, 2019, "A Tool for Extracting Text from Scanned Documents and Convert it into Editable Format", 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)

[15] M. Jha, M. Kabra, S. Jobanputra, R. Sawant, 2019, "Automation of Cheque Transaction using Deep Learning and Optical Character Recognition", International Conference, 2019.

[16] A. Ishchenko, A. G. Nesteryuk, M. V. Polyakova "The technique of extraction text areas on scanned document image using linear filtration". Applied Aspects of Information Technology, Vol. 2 No. 3. 2019.

[17] O. Petrushynskyi, Y. Kynash, O. Riznyk, N. Kustra, V. Myshchyshyn. "Smoothing the image using linear filtration". Computer Technologies of Printing. 1. pp.100-109. 2021