

Kierunek: _____ Informatyka techniczna (ITE)
Specjalność: _____ Systemy informatyki w medycynie (IMT)

PRACA DYPLOMOWA Inżynierska

**Implementacja systemu optycznego rozpoznawania
znaków oraz opracowanie środowiska eksperymentalnego
opartego o samodzielnie skonstruowany zbiór danych**

Krzysztof Zalewa

Opiekun pracy
Dr inż., Paweł Zybilewski

Słowa kluczowe: 3-6 słów kluczowych

Streszczenie

Dodaj streszczenie pracy w języku polskim. Staraj się uwzględnić wymienione na stronie tytułowej słowa kluczowe. Uwaga przedstawiony rekomendowany szablon dotyczy pracy dyplomowej pisanej w języku angielskim. W przeciwnym wypadku, student powinien samodzielnie zmienić nazwy „Chapter” na „Rodział” itp stosując odpowiednie pakiety systemu L^AT_EXoraz ustawienia w pliku *latex-settings.tex*.

Abstract

Streszczenie w języku angielskim.

Spis treści

1 Wstęp	1
1.1 Cel pracy	1
2 Przegląd literatury	3
2.1 Narzędzia OCR	3
2.1.1 Tesseract	3
2.1.2 Paddle OCR	3
2.1.3 DocTR OCR	4
2.1.4 Easy OCR	5
2.2 Zbiory danych	5
2.2.1 IAM	5
2.2.2 oldbooksdataset	6
2.3 Metryki	7
2.3.1 CER	7
2.3.2 WER	7
3 Aspekt inżynierski	9
3.1 Wybrane modyfikatory tekstu	9
3.2 Akwizycja danych	10
4 Aspekt badawczy	11
4.1 Scenariusze eksperymentów	11
4.2 Eksperymenty	12
5 Podsumowanie	13

1. Wstęp

W wyniku skanowania dokumentów powstają obrazy zawierające teksty. Jednakże większość narzędzi do przetwarzania tekstu nie jest w stanie operować na plikach graficznych. Aby rozwiązać ten problem, można zastosować algorytmy optycznego rozpoznawania znaków (Optical Character Recognition lub OCR). Są to narzędzia, które umożliwiają rozpoznanie tekstu na obrazie oraz zapisanie go w formie bardziej dogodnej do przetwarzania.

Jeden z pierwszych takich algorytmów powstał w roku 1974, a jego twórcą był Ray Kurzweil. Początkowo algorytm ten miał na celu ułatwienie funkcjonowania osobom z niepełnosprawnością wzrokową. Tekst był najpierw skanowany, a następnie przetwarzany przez OCR, dzięki czemu mógł on zostać odczytany na głos. Pierwsze algorytmy rozpoznawania znaków opierały się na skomplikowanym zbiorze zasad i reguł. Wraz ze wzrostem popularności sztucznej inteligencji oraz metod uczenia maszynowego, algorytmy OCR stopniowo zaczęły korzystać z coraz częściej korzystać z tych technologii. Dzięki temu nie tylko poprawiła się dokładność tych algorytmów, ale także zwiększyła się liczba obsługiwanych krojów pisma oraz języków. Obecnie zdecydowana większość takich algorytmów wykorzystuje zaawansowane metody sztucznej inteligencji.

Współcześnie narzędzia te nadal służą osobom z różnymi niepełnosprawnościami, jednakże zakres zastosowań algorytmów OCR znacznie się poszerzył i obejmuje inne dziedziny. Są one niezbędnym elementem w procesie digitalizacji tekstów historycznych, automatyzacji procesów biznesowych, przetwarzania dokumentów urzędowych oraz w aplikacjach mobilnych.

Mimo tego że algorytmy OCR są bardzo popularne, jest kilka kluczowych problemów, które mogą wpływać na ich funkcjonowanie. Jednym z nich jest pismo odręczne (oraz kroje pisma je imitujące). Ten typ pisma jest bardzo zróżnicowany każdy człowiek ma swój własny charakter pisma. Zapisane znaki różnią się kształtem, rozmiarem czy nachyleniem. Ponad to zdarza się, że w tym samym zdaniu te same znaki mogą się znacznie różnić. Wszystkie te czynniki sprawiają, że wiele algorytmów OCR ma mniejszą skuteczność w rozpoznawaniu pisma odręcznego od pisma maszynowego.

Kolejnym problemem jest jakość dostarczonego obrazu. W rzeczywistości rzadko spotykane są przypadki gdy, skanowany tekst jest równomiernie oświetlony, nie ma na nim cieni lub papier, na którym znajduje się tekst, jest niepogięty. Takie zniekształcenia mogą utrudnić rozdzielenie tekstu od tła, przez co skuteczność algorytmu zmalaże.

1.1. Cel pracy

Celem pracy było utworzenie autorskiego zbioru danych służącego do testowania wydajności algorytmów OCR. Zbiór zawiera dokumenty o zróżnicowanej charakterystyce, uwzględniając m.in. różne kroje, a także modyfikacje utrudniające poprawne odczytanie treści. Zbiór ten został opracowany przy pomocy interfejsu programistycznego (API) serwisu WolneLektury.pl. Serwis ten zawiera bogaty zbiór książek, opowiadań oraz wierszy, wszystkie dostępne tam

pozycje znajdują się w domenie publicznej.

2. Przegląd literatury

2.1. Narzędzia OCR

Do badań wybrano następujące cztery algorytmy optycznego rozpoznawania znaków. Badania były wykonywane na najnowszych dostępnych wersjach danego algorytmu.

2.1.1. Tesseract

Tesseract OCR to najstarszy z wybranych algorytmów optycznego rozpoznawania znaków. Został on stworzony przez firmę HP w latach 1984 - 1994. Obecnie Tesseract jest jednym z najpopularniejszych algorytmów OCR w kręgach akademickich [2, 6, 9]. Do badań użyto wersji 5.3.4 jest to znaczące gdyż wcześniejsze wersje algorytmu (do wersji 3 włącznie) nie wykorzystywały sieci neuronowych. Algorytm ten działa w następujących krokach:

1. Analiza komponentów, gdzie zarys tych komponentów jest przechowywany. Takie podejście mimo że nakłada dodatkowe koszty obliczeniowe pozwala na łatwiejsze rozpoznawanie tekstu w odwróconych kolorach (biały tekst na czarnym tle) [10].
2. Wyszukiwanie linii w komponentach. Celem tego kroku była eliminacja potrzeby korekty przekrywienia.
3. Podział linii na słowa.
4. Pierwsza iteracja rozpoznawania. Zaczynając na górze strony algorytm próbuje rozpoznać każde kolejne słowo. Jeżeli jest duże prawdopodobieństwo że słowo jest poprawne jest ono wykorzystywane do douczenia klasyfikatora. W ten sposób z każdym kolejnym słowem celność klasyfikatora powinna rosnąć.
5. Druga iteracja rozpoznawania. Po wykonaniu pierwszej iteracji jest duże prawdopodobieństwo że klasyfikator uzyskałby lepsze wyniki. Więc po raz drugi algorytm próbuje rozpoznać tekst na stronie i aktualizuje słowa które były mniej celnie rozpoznane.

2.1.2. Paddle OCR

Pomimo tego, że algorytm ten jest stosunkowo nowy (pierwsze wersje zostały wypuszczone w 2020 roku [3]). Paddle OCR jest drugim pod względem popularności algorytmem (zaraz po Tesseractie). Początkowo pierwsze wersje algorytmu skupiały się na balansie między jakością wyniku a czasem jego otrzymania. Wraz z czasem w kolejnych wersjach udoskonalano wydajność algorytmu oraz rozszerzano jego umiejętności (Np. obsługa wielu języków, rozpoznawanie pisma ręcznego). Najnowsze wersje tego algorytmu (W pracy użyto wersji

3.2.0) składają się trzech głównych modułów. **PP-OCR** Rdzeń całego algorytmu służący do rozpoznawania znaków na obrazie. **PP-Structure** Moduł służący do rozpoznawania ustrukturyzowanych obrazów (np. Zawierających tabele). **PP-ChatOCR** Moduł służący do ekstrakcji kluczowych informacji z obrazów przy pomocy dużego modelu językowego(z ang. Large Language Model lub LLM). W tej pracy zastosowany został jedynie moduł PP-OCR. Działa on w następujących krokach:

1. **Preprocesowanie** - W celu uzyskania jak najlepszej jakości obrazu algorytm może usuwać niektóre zniekształcenia oraz problemy z orientacją obrazu.
2. **Wykrycie tekstu** - Algorytm tworzy mapę prawdopodobieństwa, w której każdy piksel ma przydzieloną wartość określającą jakie jest prawdopodobieństwo, że piksel ten jest częścią obszaru tekstowego. Następnie przy pomocy binaryzacji różniczkowej (ang. Differentiable Binarization) dynamicznie określany jest próg pomiędzy tekstem a tłem. Ostatecznie na podstawie tej mapy tworzone są wielokąty będące zarysem obszaru tekstowego.
3. **Wykrycie orientacji linii** - Wykryty tekst dzielony jest na linie. Następnie algorytm upewnia się, że wykryte linie tekstu są w prawidłowej orientacji
4. **Rozpoznanie tekstu** - Poprzez zastosowanie konwolucyjnej sieci neuronowej (z ang. Convolutional Neural Network lub CNN) wykrywane są charakterystyczne elementy tekstu jak pociągnięcia, krzywe i pętle. Elementy te podawane są do rekurencyjnej sieci neuronowej (z ang. Recurrent Neural Network lub RNN), która dzięki możliwości "zapamiętania" poprzednich elementów jest w stanie odróżnić poszczególne znaki. Na koniec koneksijska klasyfikacja czasowa (z ang. Connectionist Temporal Classification lub CTC) działa jako mechanizm wyrównywania i zwijania i znajduje najbardziej prawdopodobny napis (Np. "cccczzzzzaaass" zostaje zmienione na "czas")

2.1.3. DocTR OCR

Algorytm ten jest skupiony na rozpoznawaniu dokumentów takich jak skany faktur,paragonów, formularzy czy listów. Stąd też nazwa Document Text Recognition czy docTR w skrócie. Główną filozofią tego projektu jest "bezproblemowe optyczne rozpoznanie znaków dostępne dla każdego" [8]. Algorytm ten stosuje dwuetapowe podejście do rozpoznawania tekstu:

1. **Wykrywanie tekstu** - DocTr pozwala na wykorzystanie w tym celu wielu różnych modeli. Jednakże większość z nich,podobnie jak w **Tesseract** i **Paddle OCR**, oparta jest na konwolucyjnych sieciach neuronowych. Jednakże w przeciwieństwie do tych dwóch algorytmów docTr używa piramidy cech (z ang. Feature Pyramid Network) co pozwala na odczyt tekstu w wielu różnych rozmiarach (Przydatne na przykład do odróżnienia nagłówka od adnotacji itp.). Następnie algorytm dla każdego piksela w obrazie przewiduje czy jest on w obszarze tekstowym i tworzy wielokąty wokół wykrytego tekstu.

2. **Rozpoznanie tekstu** - Podobnie jak [Paddle OCR](#) docTr najpierw rozpoznaje cechy charakterystyczne tekstu przy pomocy CNN. Jednakże do odróżnienia znaków używana jest dwukierunkowa rekurencyjna sieć neuronowa. Sieć ta różni się od zwykłej sieci RNN tym że czyta sekwencje znaków od lewej do prawej a następnie od prawej do lewej. W niektórych przypadkach ciąg liter "cl" może być bardzo zbliżony do "d". Dlatego też zabieg ten redukuje możliwość pomylenia znaków.

2.1.4. Easy OCR

EasyOCR jest jednym z nowszych algorytmów użytych w tej pracy (Pierwsza wersja pochodzi z 2019 roku [5]). Rozwojem tego projektu zajmuje się zespół Jaided AI specjalizujący się w wizji komputerowej i uczeniu maszynowym. Mimo stosunkowo krótkiej historii algorytm ten stał się popularny w środowisku deweloperów o raz w kręgach akademickich. Projekt ten priorytetyzuje prostotę, szybkość oraz wygodę w użytkowaniu. Podobnie jak [DocTR OCR](#) algorytm ten działa w dwóch krokach:

1. **Wkrywanie tekstu** - W przeciwieństwie do pozostałych algorytmów EasyOCR w tym kroku korzysta z modelu CRAFT (z ang. Character-Region Awareness For Text detection). Model ten tworzy dwie mapy prawdopodobieństwa. W pierwszej mapie zapisane jest prawdopodobieństwo tego że dany piksel znajduje się na środku znaku. Druga mapa zawiera prawdopodobieństwo tego że dany piksel jest na środku przerwy między znakami. Nakładając na siebie te dwie mapy model jest w stanie precyzyjnie wyliczyć zarys każdego znaku a następnie wyrysować wielokąt zawierający w sobie dane słowo.
2. **Rozpoznanie tekstu** - Ten krok jest już znacznie bardziej standardowy. Przebiega podobnie jak analogiczny krok w algorytmie [DocTR OCR](#).

2.2. Zbiory danych

Aby uzasadnić przydatność wykonanego zbioru danych wybrano zbiory IAM oraz old-books-dataset. Dla zbiorów tych zostały przeprowadzone badania [4, 6] w których wykonano testy dla algorytmu Tesseract.

2.2.1. IAM

IAM to zbiór ręcznie zapisanych tekstów w języku angielskim. Wykonany przez Instytut matematyki i informatyki na Uniwersytecie Breńskim. [7] Zbiór zawiera obrazy w rozdzielczości 300dpi zapisane w formacie PNG w 256 odcieniach szarości. Każdy pod katalog zawiera teksty zapisane przez jedną osobę.

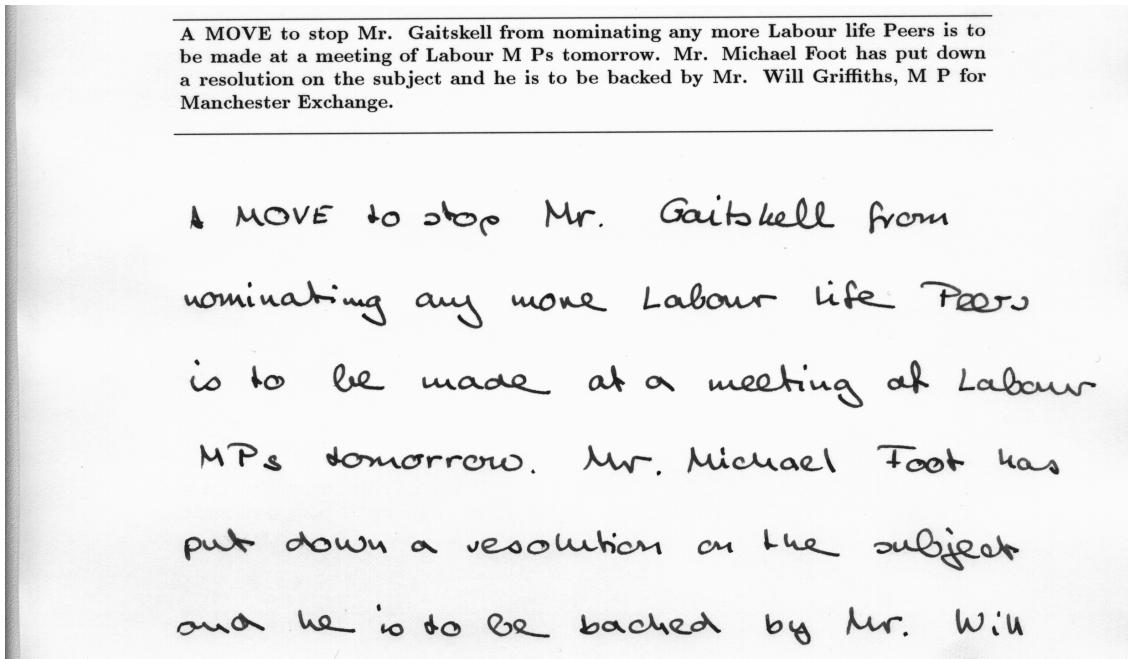


Figure 2.1: Przykładowy obraz ze zbioru danych IAM

2.2.2. oldbooksdataset

Zbiór udostępniony na platformie git hub zawierający skany książek w języku angielski. Książki zapisane są w formacie .tiff w rozdzielczości 300dpi oraz 500dpi [1].

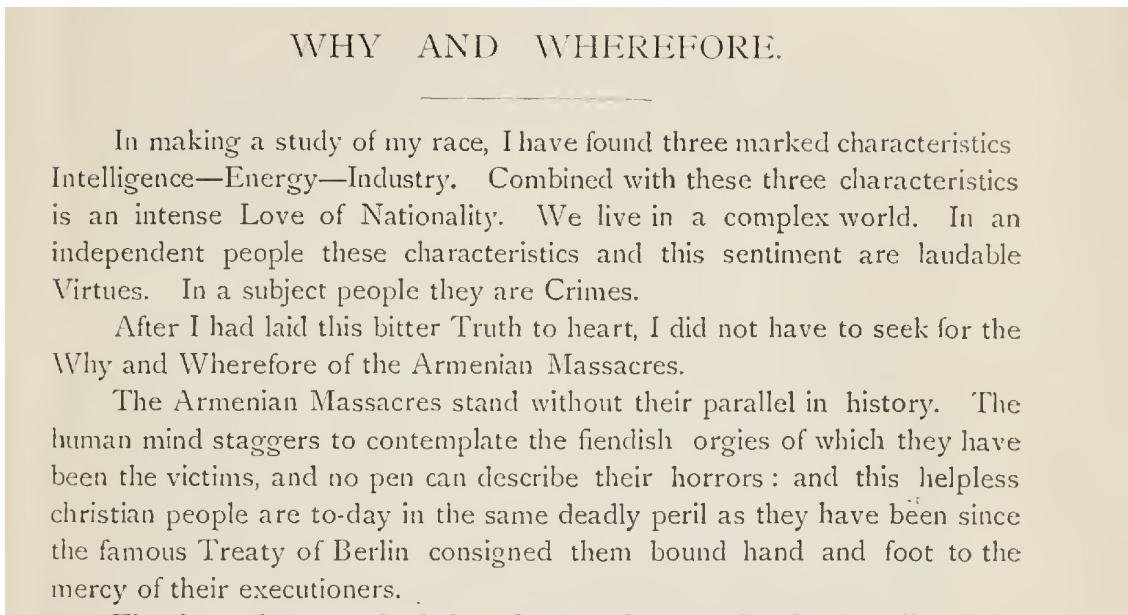


Figure 2.2: Przykładowy obraz ze zbioru danych old-books-dataset

2.3. Metryki

Do oceny wyników zastosowano dwie główne metryki. **CER** i **WER** zasadniczą różnicą między nimi jest celność porównań. CER porównuje na poziomie pojedynczych znaków natomiast, z kolei WER porównuje na poziomie poszczególnych słów. Z tego powodu CER jest przydatniejsze w kontekstach w których nawet pojedynczy znak może zmienić znaczenie słowa (np. w medycynie). Natomiast WER jest lepsze do porównywania spójności zdań itp.

2.3.1. CER

CER (Character Error Rate z ang. częstotliwość błędnych znaków) to metryka dzięki której możliwa jest ocena różnic między tekstem wytworzonym poprzez model OCR a tekstem rzeczywistym. W tym przypadku CER obliczane jest poprzez zsumowanie operacji(wstawień, usunięć oraz zamian znaków) potrzebnych do uzyskania tekstu rzeczywistego.

$$CER = \frac{S + D + I}{N_c}$$

Gdzie:

- S - Liczba zamian znaków (ang. Substitutions)
- D - Liczba usunięć znaków (ang. Deletions)
- I - Liczba wstawień znaków (ang. Inserts)
- N_c - Liczba znaków w tekście (ang. Number of characters)

Na przykład

Tekst oryginalny: Życiem wschód, śmierci południe;

Tekst wygenerowany przez model: Zyciem wschod, siercia poudniex;

Aby przekształcić tekst wygenerowany do tekstu oryginalnego należy wykonać 4 zamiany (Brakujące znaki polskie), 1 wstawienie (Brakujące 'm' w tekście wygenerowanym) oraz 1 usunięcie ('x' nie występuje w tekście oryginalnym). Więc $CER = 6/28 = 0.2141 \approx 21.4\%$

2.3.2. WER

WER (Word Error Rate z ang. częstotliwość błędnych słów) podobnie jak CER jest to metryka dzięki której możliwa jest ocena różnic między tekstem wytworzonym poprzez model OCR a tekstem rzeczywistym. Jak sama nazwa wskazuje WER porównuje tekst na poziomie poszczególnych słów.

$$WER = \frac{S + D + I}{N_w}$$

Gdzie:

- S - Liczba zamian słów (ang. Substitutions), czyli słowa które występują w obu tekstach ale te w tekście są różne od tych w tekście orginalnym.
- D - Liczba usunięć słów (ang. Deletions), czyli słowa które występują w tekście orginalnym jednakże nie ma ich w tekście wygenerowanym.
- I - Liczba wstawień słów (ang. Inserts), czyli słowa nadmiarowe których nie ma w tekście orginalnym.
- N_w - Liczba słów w tekście (ang. Number of words)

Na przykład

Tekst orginalny: Życiem wschód, śmierci południe;

Tekst wygenerowany przez model: Zyciem wschod, siercia poudniex;

Aby przekształcić tekst wygenerowany do tekstu oryginalnego należy wykonać 4 zamiany (Słowa zbliżone do oryginału ale nie takie same). Więc WER=4/4=1=100%

3. Aspekt inżynierski

3.1. Wybrane modyfikatory tekstu

Kroje pisma

Do wykonania autorskiego zboiru danych wykorzystano następujące kroje pisma.

Allura Regular

Ten krój ma symulować pismo odręczne. Jednakże jak widać na poniższej próbce (Rys 3.1) znaki w tym kroju są ze sobą połączone. Może to utrudniać rozpoznanie poszczególnych znaków, co wpłynie na wyniki algorytmów.



Lorem ipsum dolor sit amet

Figure 3.1

Caveat

Podobnie jak poprzedni krój, Caveat ma przypominać pismo odręczne. Jednakże w przeciwieństwie do poprzednika w tym kroju znaki są wraźnie rozdzielone (Rys 3.2).



Lorem ipsum dolor sit amet

Figure 3.2

Times New Roman

W przeciwieństwie do poprzednich krojów Times New Roman nie jest stylizowany na pismo odręczne (Rys 3.3). Jest to jeden z bardziej popularnych krojów szeryfowych. Szeryfy to poprzeczne lub ukośne zakończenia głównych pociągnięć znaków.

Łorem ipsum dolor sit amet
Łorem ipsum dolor sit amet

Figure 3.3

Ubuntu

Kroje szeryfowe są popularne w publikacjach papierowych, jednakże podczas czytania tekstu elektronicznego mogą być one męczące dla oczu. Dlatego obecnie coraz częściej stosowane są kroje bez-szeryfowe (ang. sans-serif). Ubuntu jest przykładem takiego kroju (Rys 3.4).

Łorem ipsum dolor sit amet

Figure 3.4

Znieształcenia obrazu

1.

3.2. Akwizycja danych

Interfejs programistyczny [WolneLektury.pl](#)
Ekstrakcja tekstu
Przygotowanie obrazów

4. Aspekt badawczy

Protokół eksperymentalny W celu zapewnienia rzetelności procedury badawczej, wszystkie eksperymenty zostały przeprowadzone przy pomocy 2-foldowej walidacji krzyżowej powtórzonej 5-krotnie. Otrzymane w ten sposób wyniki porównano przy pomocy testu Wilcoxon'a dla par obserwacji gdzie $\alpha = 0.05$. Aby udowodnić że różnice między wynikami nie są losowe użyto połączonego testu F dla walidacji krzyżowej (Combined 5×2 CV F-test).
Metryki oceny Skuteczność algorytmów oceniano na podstawie współczynnika błędów znakowych (Character Error Rate). Dla oceny wydajności obliczeniowej wykorzystano średni czas przetwarzania pojedynczej strony tekstu.

Pytania badawcze

1. Który algorytm zapewnia najlepszy kompromis między dokładnością a szybkością przetwarzania?
2. W jakim stopniu opracowany autorski zbiór danych jest porównywalny z uznanymi benchmarkami OCR pod względem trudności i zdolności do weryfikacji podstawowej skuteczności algorytmów?
3. W jakim stopniu rodzaj i stopień wybranych znieksztalceń wpływa na wyniki algorytmów OCR?
4. W jakim stopniu wybrane kroje pisma wpływają na dokładność działania wybranych algorytmów?

4.1. Scenariusze eksperymentów

Eksperiment 1 Porównanie szybkości przetwarzania a jakości wyników Celem pierwszego eksperymentu jest analiza związku między czasem przetwarzania a jakością otrzymanych wyników. Eksperiment zostanie przeprowadzony na główny podzbiorze(Bez znieksztalceń oraz standardowy krój pisma) autorskiego zbioru danych.

Eksperiment 2 Porównanie autorskiego zbioru z uznanymi benchmarkami Celem tego eksperymentu jest uzasadnienie, autorski zbiór danych jest porównywalny z innymi benchmarkami algorytmów OCR(Zbiory IAM oraz old-books-dataset). Podobnie jak w przypadku eksperymentu 1 w tym eksperymencie zostanie wykorzystany główny podzbior autorskiego zbioru.

Eksperiment 3 Wpływ znieksztalceń obrazu Trzeci eksperiment ma na celu zbadanie wpływu zastosowanych znieksztalceń obrazu (Pochylenie perspektywiczne, falowanie tekstu, częściowe zacienienie, Symulacja kropli wody na obiektywie) na otrzymane wyniki. Dla każdego rodzaju znieksztalconia zostanie wygenerowany podzbior. Eksperiment zostanie przeprowadzony na tych podzbiorach.

Eksperyment 4 Wpływ krojów pisma Natomiast celem czwartego eksperymentu jest zbadanie wpływu wybranych krojów pisma (Brightlast, Tangerine,Ubuntu-B,Times New Roman oraz Fira Mono) na otrzymane wyniki. Dla każdego kroju pisma zostanie wygenerowany podzbiór. Eksperyment zostanie przeprowadzony na tych podzbiorach.

4.2. Eksperymenty

5. Podsumowanie

Bibliografia

- [1] P. Barcha. Old books dataset. <https://github.com/PedroBarcha/old-books-dataset>, 2024. Accessed: 2024.
- [2] A. Chowdhury, A. A. Sami, S. M. P. Mamun, S. Absar, F. Biswas, and M. Kohinoor. Performance analysis of tesseract and easyocr for bangla optical character recognition on the novel bangla crosshair dataset. In *Proceedings of the International Conference on Computer and Communication Systems*, 01 2025.
- [3] C. Cui, T. Sun, M. Lin, T. Gao, Y. Zhang, J. Liu, X. Wang, Z. Zhang, C. Zhou, H. Liu, Y. Zhang, W. Lv, K. Huang, Y. Zhang, J. Zhang, J. Zhang, Y. Liu, D. Yu, and Y. Ma. Paddleocr 3.0 technical report. *arXiv preprint*, 2025.
- [4] T. Hegghammer. Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, 5:861–882, 2022.
- [5] JaidedAI. Ready-to-use ocr. <https://github.com/JaidedAI/EasyOCR>, 2024.
- [6] Y. Li. Synergizing optical character recognition: A comparative analysis and integration of tesseract, keras, paddle, and azure ocr. *University of Sydney Technical Reports*, 45:45–60, 2023.
- [7] U. Marti and H. Bunke. The iam-database: An english sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, 2002.
- [8] Mindee. doctr: Document text recognition. <https://github.com/mindee/doctr>, 2021.
- [9] V. S and S. A. Performance comparison of ocr tools. *International Journal of UbiComp*, 6(3):19–30, July 2015.
- [10] R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633, 2007.

Spis ilustracji

2.1	Przykładowy obraz ze zbioru danych IAM	6
2.2	Przykładowy obraz ze zbioru danych old-books-dataset	6
3.1		9
3.2		9
3.3		10
3.4		10

Spis tabel