

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 Metryka CER	7
2.4 Szum Perlina	7
2.5	8
3 Modyfikatorem tekstu	9
3.1 Kroje pisma	9
3.2 Zniekształcenia obrazu	10
4 Akwizycja danych	17
4.1 Interfejs programistyczny WolneLektury.pl	17
4.2 Ekstrakcja tekstu	21
4.3 Przygotowanie obrazów	22
5 Aspekt badawczy	25
5.1 Scenariusze eksperymentów	25
6 Wyniki	27
7 Podsumowanie	29

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.

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 pisma, 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, 10]. 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) [11].
2. Wyszukiwanie linii w komponentach. Celem tego kroku była eliminacja potrzeby korekty przekrystwienia.
3. Podział linii na słowa.
4. Pierwsza iteracja rozpoznawania. Zaczynając na górze strony algorytm próbuje rozpoznać każde kolejne słowo. Tokeny rozpoznane z wysokim stopniem pewności są wykorzystywane do douczenia klasyfikatora. W ten sposób z każdym kolejnym słowem celność klasyfikatora powinna rosnąć.
5. Druga iteracja rozpoznawania. Po wykonaniu pierwszej iteracji istnieje szansa, 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. obsługą 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 koneksjonistyczna klasyfikacja czasowa (z ang. Connectionist Temporal Classification lub CTC) działa jako mechanizm wyrównywania i zwijania i znajduje najbardziej prawdopodobny napis (Np. "cccczzzzaaass" 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 rozpoznawanie 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 śród 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:

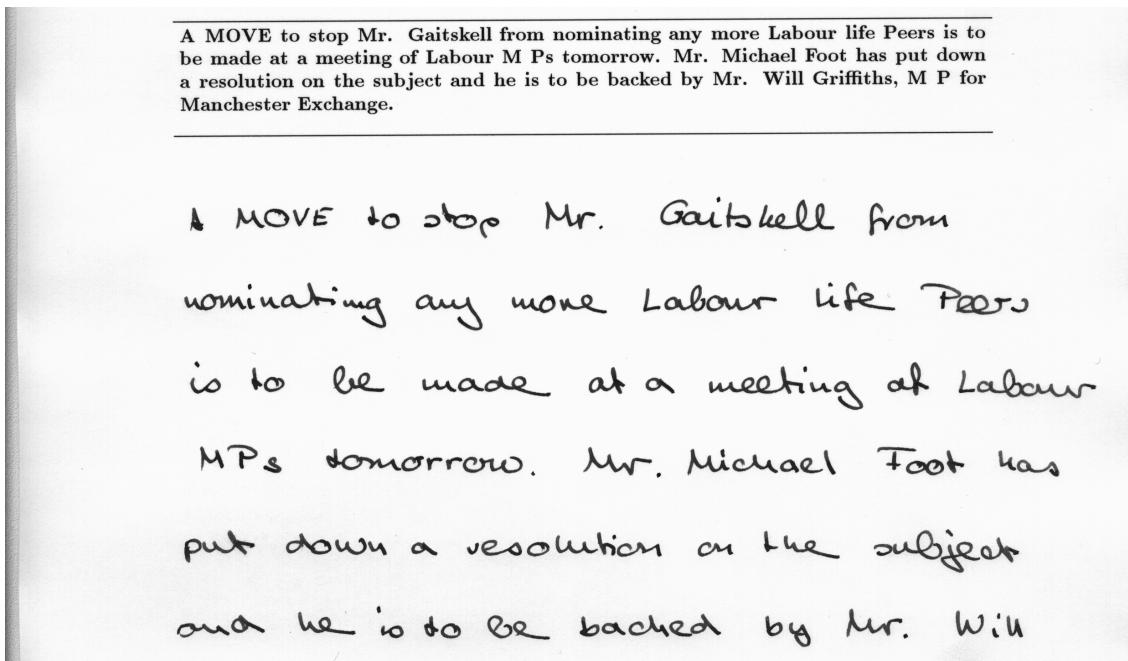
1. **Wkrywanie tekstu** - W przeciwieństwie do pozostałych algorytmów EasyOCR w tym kroku korzysta z modelu CRAFT (z ang. Character-Region Awareness For Text detection). Model ten tworzy dwie mapy prawdopodobieństwa. W pierwszej mapie zapisane jest prawdopodobieństwo że dany piksel znajduje się na środku 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

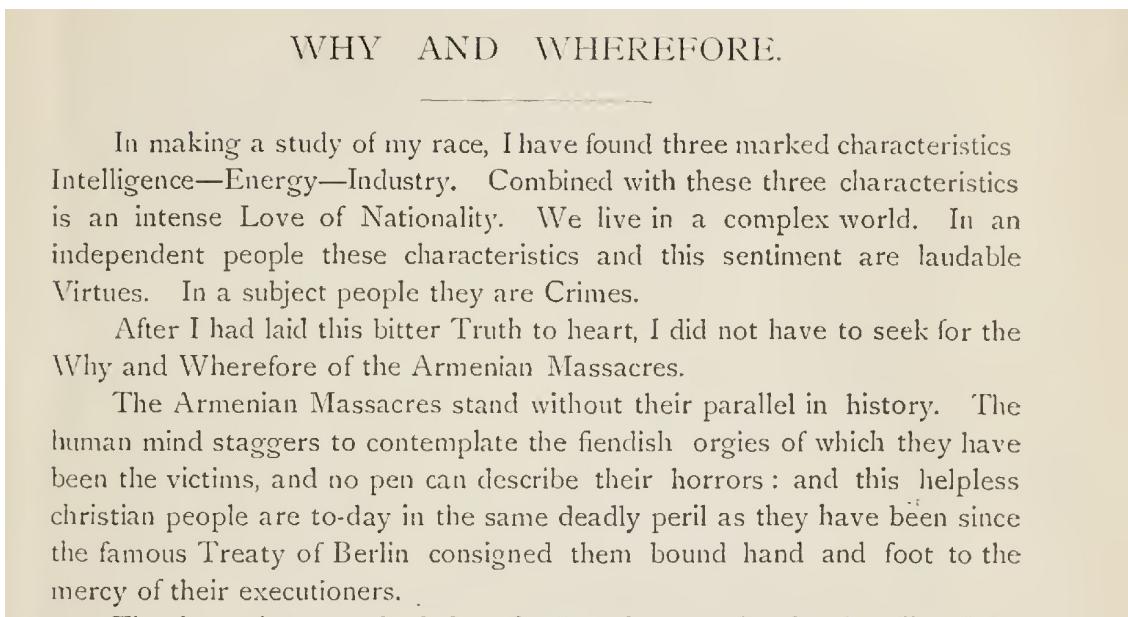
IAM to zbiór ręcznie zapisanych tekstów w języku angielskim. Wykonany przez Instytut matematyki i informatyki na Uniwersytecie Breńskim. [7] Zbiór zawiera obrazy w rozdzielcości 300dpi zapisane w formacie PNG w 256 odcieniach szarości. Każdy pod katalog zawiera teksty zapisane przez jedną osobę.



Rysunek 2.1: Przykładowy obraz ze zbioru danych IAM

2.2.2. oldbooksdataset

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



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

2.3. Metryka 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: Źyciem 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 $CER = 6/28 = 0.2141 \approx 21.4\%$

2.4. Szum Perlina

Ten typ generacji pseudo-losowego szumu został zaproponowany przez Kena Perlina w 1985 [9]. W przeciwieństwie do czysto losowych danych, które charakteryzują się ostrymi, nieciągłymi zmianami, szum Perlina generuje wartości zmieniające się w sposób płynny i gradualny, co lepiej odzworowuje procesy obserwowane w środowisku naturalnym.

Algorytm, choć przedstawiony na Rysunku () na przykładzie jednowymiarowym, może być rozszerzony do dwóch lub trzech wymiarów. Wielowymiarowe warianty szumu Perlina znajdują zastosowanie w wielu środowiskach dla przykładu w grafice komputerowej w szczególności podczas proceduralnej generacji tekstur. Z szumu Perlina korzysta się też w produkcji filmów korzystających z grafiki generowanej komputerowo (CGI) oraz tworzeniu środowisk w grach komputerowych(np. w grze Minecraft proces generacja terenu korzysta z tego algorytmu).

Implementacja szumu Perlina przebiega w trzech głównych etapach.

Definicja przestrzeni

W pierwszym etapie generowana jest regularna siatka w przestrzeni n -wymiarowej, gdzie każdemu węźlowi siatki przypisywany jest losowy wektor jednostkowy, określający gradient w danym punkcie przestrzeni.

Obliczenie iloczynu skalarnego

Dla dowolnego punktu P w przestrzeni identyfikowana jest komórka siatki, w której się on znajduje. Następnie dla każdego wierzchołka V_i tej komórki obliczany jest wektor przesunięcia

$\vec{d}_i = P - V_i$ oraz iloczyn skalarny między tym wektorem a wektorem gradientu przypisany do wierzchołka:

$$s_i = \vec{d}_i \cdot \vec{g}_i$$

Interpolacja

Ostateczna wartość szumu w punkcie P obliczana jest poprzez interpolację wartości s_i uzyskanych dla wszystkich wierzchołków komórki. Interpolacja przeprowadzana jest przy użyciu funkcji wygładzającej, która zapewnia ciągłość pochodnych na granicach komórek.

W wykorzystanej implementacji, która pochodzi z biblioteki `noise` podczas generacji szumu Perlina można podać następujące parametry.

1. **x,y** - Kordynaty generowanego punktu.
2. **Liczba oktaw (Octaves)** - Określa ilość nakładanych warstw szumu. Większa liczba oktaw powoduje dodanie detali o wyższych częstotliwościach, co skutkuje bardziej złożonym i szczegółowym wzorem wyjściowym.
3. **Trwałość (Persistence)** - Współczynnik określający amplitudę kolejnych oktaw. Przyjmuje wartości z zakresu $[0, 1]$. Wysokie wartości powodują zachowanie większej ilości szczegółów, generując ostrzejsze wzory. Natomiast niskie wartości bardziej wygładzonych rezultatów.
4. **Lacunarity** - Parametr określający współczynnik zmiany częstotliwości między kolejnymi oktawami. Wartość musi być większa od 1. Wyższe wartości powodują szybszy wzrost częstotliwości w kolejnych oktawach, co skutkuje większym zróżnicowaniem skali detali.
5. **Repeatax** - Parametr określający powtarzalność szumu w osi X. Jeżeli parametr ten równa się szerokości obrazu wzór nie będzie się powtarzał.
6. **Repeaty** - Parametr określający powtarzalność szumu w osi Y. Jeżeli parametr ten równa się wysokości obrazu wzór nie będzie się powtarzał.

2.5.

3. Modyfikatorem tekstu

3.1. Kroje pisma

Jednym z problemów, na które często napotykają się algorytmy OCR, 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. Ponadto 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.

W tej pracy wybrano cztery kroje pisma. Pierwsze dwa mają przypominać pismo odręczne, pozostałe dwa są popularnymi przykładami standardowych krojów.

Allura Regular

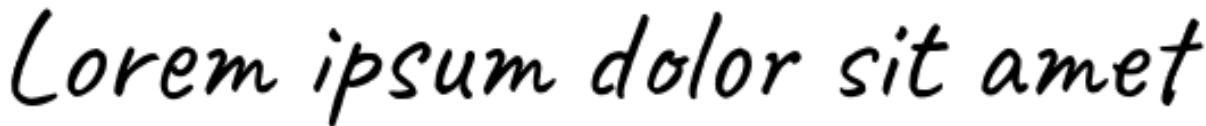
Ten krój ma symulować pismo odręczne. Jednakże jak widać na poniższej próbce (Rys 3.1) znaki w tym kroju są ze sobą połączone. Może to utrudniać rozpoznanie poszczególnych znaków, co wpłynie na wyniki algorytmów.

A sample of the Allura Regular font showing the text "Lorem ipsum dolor sit amet". The font has a cursive, handwritten style with connected letters.

Rysunek 3.1: Próbka tekstu z użyciem kroju Allura ("Lorem ipsum dolor sit amet")

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).

A sample of the Caveat font showing the text "Lorem ipsum dolor sit amet". The font has a cursive, handwritten style with clearly separated letters.

Rysunek 3.2: Próbka tekstu z użyciem kroju Caveat ("Lorem ipsum dolor sit amet")

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 szerygowych. Szeryfy to poprzeczne lub ukośne zakończenia głównych pociągnięć znaków.

Rysunek 3.3: Próbka tekstu z użyciem kroju Times New Roman ("Lorem ipsum dolor sit amet")

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).



Rysunek 3.4

Próbka tekstu z użyciem kroju Ubuntu ("Lorem ipsum dolor sit amet")

3.2. Zniekształcenia obrazu

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 zmaleje.

W tej pracy wybrano trzy rodzaje takich zniekształceń obrazu. Są to:

Przesunięcie perspektywy

Obrazy poddawane są losowej deformacji przestrzennej poprzez przekształcenie projektowe. Dla każdego obrazu generowane są cztery współczynniki deformacji z zakresu [0.01, 0.2], które określają przemieszczenie narożników. Na podstawie punktów źródłowych i docelowych obliczana jest macierz transformacji, a następnie aplikowana na obraz wejściowy przy użyciu funkcji `cv2.warpPerspective`.

Kod źródłowy 3.1: Fragment autorskiej klasy ImageGenerator odpowiedzialnej za generację przekrzywionych obrazów

```
1 import os
2 import random
3
4 import cv2
5 import numpy as np
6
7
8 class ImageGenerator:
9     def tilt_img(self, input_dir: str, file_name: str):
10         os.makedirs(self.TILT_DIR, exist_ok=True)
11         input_path = os.path.join(input_dir, file_name)
```

```
12     output_path = os.path.join(self.TILT_DIR, file_name)
13
14     original_image = cv2.imread(input_path)
15     height, width = original_image.shape[:2]
16
17     src_points = np.float32([[0, 0], [width, 0], [0, height], [
18         width, height]])
19
20     random.seed(self.seed)
21     tilt_tl: float = random.uniform(0.01, 0.2)
22     tilt_tr: float = random.uniform(0.01, 0.2)
23     tilt_bl: float = random.uniform(0.01, 0.2)
24     tilt_br: float = random.uniform(0.01, 0.2)
25
26     dst_points = np.float32(
27         [
28             [width * tilt_tl, 0],
29             [width * (1 - tilt_tr), 0],
30             [0, height * (1 - tilt_bl)],
31             [width * (1 - tilt_br), height * (1 - tilt_br)],
32         ]
33     )
34
35     matrix = cv2.getPerspectiveTransform(src_points, dst_points)
36
37     tilted_image = cv2.warpPerspective(original_image, matrix, (
38         width, height))
39     cv2.imwrite(output_path, tilted_image)
```

Pasterz jakiś z tych stron rodem.
Smutny był bardzo, ale przyczyny
Smutku nie mówił nikomu,
Odbił się potem od swej drużyny
I nie powrócił do domu.
Często widziałem, czy świecą zorza,
Czyli księżyc w pełnym blasku,

Rysunek 3.5: Fragment tekstu po nałożeniu efektu zmiętej kartki

Zacienienie

Na oryginalny obraz nałożono losowo wygenerowane cienie, wygenerowane przy pomocy metody `pnoise2` z biblioteki `noise`, która służy do generowania losowego szumu Perlina.

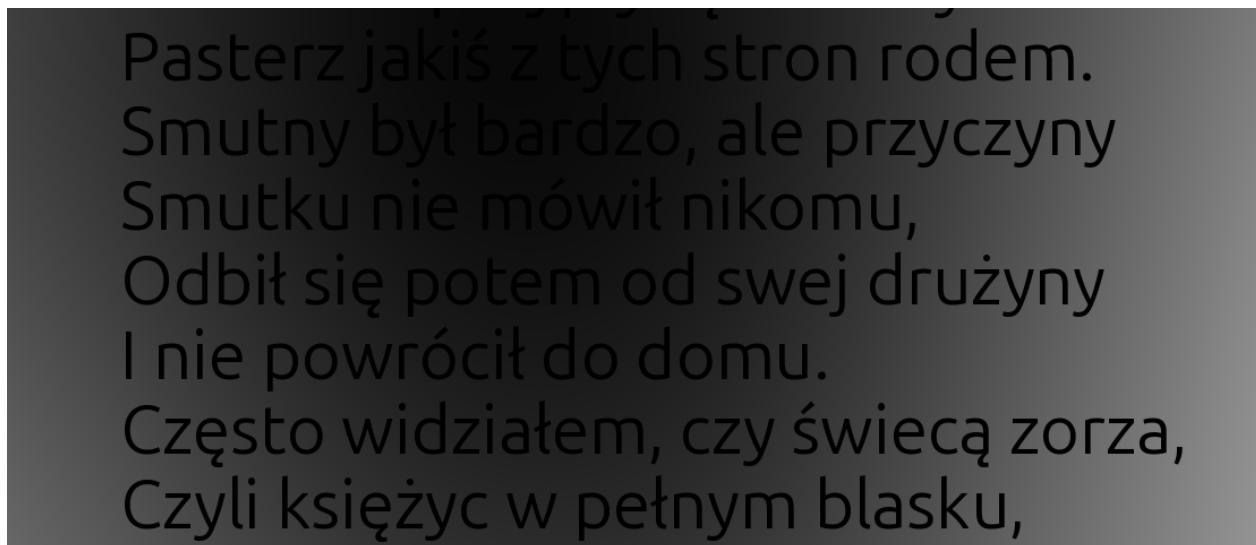
Dobór wartości parametrów tej funkcji przeprowadzono metodą eksperymentalną, poprzez wielokrotną iterację z różnymi konfiguracjami, aż do osiągnięcia pożądanego efektu w postaci delikatnego, lokalnego zacienienia.

Kod źródłowy 3.2: Fragment autorskiej klasy ImageGenerator odpowiedzialnej za generację zacienionych obrazów

```
1 import os
2
3 import cv2
4 import numpy as np
5 from noise import pnoise2
6
7
8 class ImageGenerator:
9     def shadow_img(self, input_dir: str, file_name: str):
10         scale = 5000.0
11         octaves = 2
12         persistence = 0.5
13         lacunarity = 2.0
14
15         os.makedirs(self.SHADOW_DIR, exist_ok=True)
16
17         input_path = os.path.join(input_dir, file_name)
18         output_path = os.path.join(self.SHADOW_DIR, file_name)
19
20         shadow = np.zeros((self.A4_HEIGHT, self.A4_WIDTH))
21
22         for i in range(self.A4_HEIGHT):
23             for j in range(self.A4_WIDTH):
24                 shadow[i][j] = pnoise2(
25                     i / scale,
26                     j / scale,
27                     octaves=octaves,
28                     persistence=persistence,
29                     lacunarity=lacunarity,
30                     repeatx=self.A4_WIDTH,
31                     repeaty=self.A4_HEIGHT,
32                     base=self.seed,
33                 )
34
35         normalized_shadow = (shadow - shadow.min()) / (shadow.max()
36                         - shadow.min())
37
38         shadow_dark = normalized_shadow + 0.05
39         original_image = cv2.imread(input_path)
40
41         shadow_rgb = cv2.cvtColor(shadow_dark.astype(np.float32),
42                               cv2.COLOR_GRAY2BGR)
43         image_with_shadow = original_image.astype(np.float32) *
44             shadow_rgb
45         image_with_shadow = np.clip(image_with_shadow, 0, 255).
46             astype(np.uint8)
```

```
44 cv2.imwrite(output_path, image_with_shadow)
```

Wygenerowany w ten sposób szum został poddany normalizacji. Następnie otrzymaną mapę cieni nałożono na oryginalny obraz.



Rysunek 3.6: Fragment tekstu po zacienieniu

Efekt zmiętej kartki

Oryginalny obraz został poddany deformacji, obliczonej na podstawie losowo wygenerowanego szumu Perlina. W celu otrzymania tego efektu, podobnie jak w przypadku obrazów zacienionych skorzystano z metody `pnoise2` biblioteki `noise`. Także procedura wyboru parametrów była zbliżona do poprzednika. Jednakże w tym przypadku pożądanym efektem były podłużne, nieregularne formy imitujące naturalne zgięcia i deformacje materiału papieru, takie jak występują w fizycznych, użytkowanych dokumentach.

Kod źródłowy 3.3: Fragment autorskiej klasy `ImageGenerator` odpowiedzialnej za generację zmiętych obrazów

```
1 import os
2
3 import cv2
4 import numpy as np
5 from noise import pnoise2
6
7
8 class ImageGenerator:
9     def wrinkle_img(self, input_dir: str, file_name: str):
10         scale = 500.0
11         octaves = 3
12         persistence = 0.5
13         lacunarity = 2.0
14
15         os.makedirs(self.WRAPPER_DIR, exist_ok=True)
16
17         input_path = os.path.join(input_dir, file_name)
```

```
18     output_path = os.path.join(self.WRINKLE_DIR, file_name)
19
20     wrinkles = np.zeros((self.A4_HEIGHT, self.A4_WIDTH))
21
22     for i in range(self.A4_HEIGHT):
23         for j in range(self.A4_WIDTH):
24             wrinkles[i][j] = pnoise2(
25                 i / scale,
26                 j / scale,
27                 octaves=octaves,
28                 persistence=persistence,
29                 lacunarity=lacunarity,
30                 repeatx=self.A4_WIDTH,
31                 repeaty=self.A4_HEIGHT,
32                 base=self.seed,
33             )
34
35     normalized_wrinkles = (wrinkles - wrinkles.min()) / (
36         wrinkles.max() - wrinkles.min()
37     )
38
39     original_image = cv2.imread(input_path)
40
41     scale_intensity = 50
42     disp_x = (normalized_wrinkles - 0.5) * scale_intensity
43     disp_y = (normalized_wrinkles - 0.5) * scale_intensity
44
45     x, y = np.meshgrid(np.arange(self.A4_WIDTH), np.arange(self.
46         A4_HEIGHT))
47
48     map_x = (x + disp_x).astype(np.float32)
49     map_y = (y + disp_y).astype(np.float32)
50
51     wrinkled_image = cv2.remap(original_image, map_x, map_y, cv2
52         .INTER_LINEAR)
53     cv2.imwrite(output_path, wrinkled_image)
```

Po wygenerowaniu otrzymany szum był normalizowany, w celu otrzymania mapy przesunięć. Następnie otrzymaną w ten sposób mapę nakładano na oryginalny obraz przy pomocy wbudowanych funkcji biblioteki cv2.



Królewiec z
Wtenczas przypłynął z Litwy
Pasterz jakiś z tych stron rodem.
Smutny był bardzo, ale przyczyny
Smutku nie mówił nikomu,
Odbił się potem od swej drużyny
I nie powrócił do domu,
Często widziałem, czy świecą zorza,
Często w pełnym blasku,
Często w pełnym blasku, morza

Rysunek 3.7: Fragment tekstu po deformacji przestrzennej

4. Akwizycja danych

4.1. Interfejs programistyczny WolneLektury.pl

Dane do zbioru zostały pozyskane z serwisu WolneLektury.pl. Dostęp do tych danych możliwy jest poprzez publiczny interfejs programistyczny (Application Programming Interface – API), dostępny pod adresem bazowym: <https://wolnelektury.pl/api/>.

Architektura API opiera się na zasadach REST (Representational State Transfer), oferując strukturalny dostęp do zasobów poprzez logiczne endpointy. Serwis WolneLektury.pl udostępnia wiele punktów dostępu, jednakże te najważniejsze dla tej pracy to:

- /api/books/ – Wszystkie utwory
- /api/authors/ – Lista autorów

Domyślnie dane otrzymywane w wyniku zapytania są serializowane w formacie JSON, możliwa jest też zmiana na format XML (poprzez dodanie parametru ?format=xml do zapytania). Jednakże w tej pracy wykorzystano domyślny format.

Mechanizm filtrowania i precyzyjnego wyszukiwania rekordów realizowany jest poprzez rozbudowę ścieżki URL o odpowiednie parametry. Na przykład zapytanie pod adres /api/authors/adam-mickiewicz/ zwraca informacje o autorze "adam-mickiewicz".

```
{  
    "name": "Adam Mickiewicz",  
    "url": "https://wolnelektury.pl/katalog/autor/adam-mickiewicz/",  
    "sort_key": "mickiewicz adam",  
    "description": "<dl><dt>Ur.</dt><dd> 24 grudnia 1798 r. w Zaosiu ko",  
    "description_pl": "<dl><dt>Ur.</dt><dd> 24 grudnia 1798 r. w Zaosiu  
    "plural": "",  
    "genre_epoch_specific": false,  
    "adjective_feminine_singular": "",  
    "adjective_nonmasculine_plural": "",  
    "genitive": "Adama Mickiewicza",  
    "collective_noun": ""  
}
```

Rysunek 4.1

Najważniejszą funkcjonalnością tego API z punktu widzenia tej pracy jest możliwość łączenia endpointów w łańcuchy zapytań. Dzięki temu możliwe jest dokładne definiowanie podzbiorów danych. Dla ilustracji, zapytanie /api/authors/adam-mickiewicz/books/ zwraca metadane dla książek, których autorem jest Adam Mickiewicz.

```
[  
  {  
    "kind": "Liryka",  
    "full_sort_key": "mickiewicz adam$ballady i romanse$291",  
    "title": "Ballady i romanse",  
    "url": "https://wolnelektury.pl/katalog/lektura/ballady-i-romanse/",  
    "cover_color": "#db4b16",  
    "author": "Adam Mickiewicz",  
    "cover": "book/cover/ballady-i-romanse.jpg",  
    "epoch": "Romantyzm",  
    "href": "https://wolnelektury.pl/api/books/ballady-i-romanse/",  
    "has_audio": true,  
    "genre": "Ballada",  
    "simple_thumb": "https://wolnelektury.pl/media/book/cover_api_thumb/ballady-i-romanse_BZcruYT.jpg",  
    "slug": "ballady-i-romanse",  
    "cover_thumb": "book/cover_thumb/ballady-i-romanse_dr22a1D.jpg",  
    "liked": null  
  },  
]
```

Rysunek 4.2

Wynik takiego zapytania zawiera listę skróconych opisów pozycji. Aby otrzymać dokładniejsze dane na temat danej książki należy wykorzystać atrybut "href", który zawiera link do szczegółowego opisu pozycji.

```
{  
  "title": "Ballady i romanse",  
  "url": "https://wolnelektury.pl/katalog/lektura/ballady-i-romanse/",  
  "language": "pol",  
  "epochs": [  
    {  
      "url": "https://wolnelektury.pl/katalog/epoka/romantyzm/",  
      "href": "https://wolnelektury.pl/api/epochs/romantyzm/",  
      "name": "Romantyzm",  
      "slug": "romantyzm"  
    }  
  ],  
  "genres": [  
    {  
      "url": "https://wolnelektury.pl/katalog/gatunek/ballada/",  
      "href": "https://wolnelektury.pl/api/genres/ballada/",  
      "name": "Ballada",  
      "slug": "ballada"  
    }  
  ],  
  "kinds": [  
    {  
      "url": "https://wolnelektury.pl/katalog/rodzaj/liryka/",  
      "href": "https://wolnelektury.pl/api/kinds/liryka/",  
      "name": "Liryka",  
      "slug": "liryka"  
    }  
  ],  
  "authors": [  
    {  
      "url": "https://wolnelektury.pl/katalog/autor/adam-mickiewicz/",  
      "href": "https://wolnelektury.pl/api/authors/adam-mickiewicz/",  
      "name": "Adam Mickiewicz",  
      "slug": "adam-mickiewicz"  
    }  
  ],  
  "translators": [],  
  "fragment_data": {  
    "title": "Adam Mickiewicz, Ballady i romanse, To lubię",  
    "html": "<div class=\"stanza\"><div class=\"verse verse-indent-1\">Raz, gdy do Ruty jadę w czas noclegu,</div>\n</div>\n"
```

Rysunek 4.3

```

    "parent": null,
    "preview": false,
    "epub": "https://wolnelektury.pl/media/book epub/ballady-i-romanse.epub",
    "mobi": "https://wolnelektury.pl/media/book/mobi/ballady-i-romanse.mobi",
    "pdf": "https://wolnelektury.pl/media/book/pdf/ballady-i-romanse.pdf",
    "html": "https://wolnelektury.pl/media/book/html/ballady-i-romanse.html",
    "txt": "https://wolnelektury.pl/media/book/txt/ballady-i-romanse.txt",
    "fb2": "https://wolnelektury.pl/media/book/fb2/ballady-i-romanse.fb2",
    "xml": "https://wolnelektury.pl/media/book/xml/ballady-i-romanse.xml",
    "media": [],
    "audio_length": "",
    "cover_color": "#db4b16",
    "simple_cover": "https://wolnelektury.pl/media/book/cover_simple/ballady-i-romanse_P6innIl.jpg",
    "cover_thumb": "https://wolnelektury.pl/media/cache/65/dd/65dd19f39a51dd3c10dc6c3170322e21.jpg",
    "cover": "https://wolnelektury.pl/media/book/cover/ballady-i-romanse.jpg",
    "simple_thumb": "https://wolnelektury.pl/media/book/cover_api_thumb/ballady-i-romanse_BZcruYT.jpg",
    "isbn_pdf": "ISBN 978-83-288-0533-0",
    "isbn_epub": "ISBN 978-83-288-3529-0",
    "isbn_mobi": "ISBN 978-83-288-4615-9"
}

```

Rysunek 4.4

W szczegółowych danych na temat książki najważniejsze, z punktu widzenia tej pracy, były pola "epub", "html", "txt" oraz "xml". Zawierają one linki do pobrania zawartości książki w odpowiednim formacie.

Kod źródłowy 4.1: Fragment kodu odpowiedzialny za pobieranie metadanych pozycji.

```

1 import json
2 import os
3
4 import requests
5 from utils.path import is_valid
6
7
8 def get_book_list(use_cache: bool) -> None | list[Book]:
9     os.makedirs(CACHE_DIR, exist_ok=True)
10    path = os.path.join(CACHE_DIR, f"{AUTHOR}.json")
11    if is_valid(path) and use_cache:
12        print(f"File {AUTHOR}.json already exists, skipping url call
13            ")
14        return open_json(path)
15
16    try:
17        print(f"Getting: {url}")
18        response = requests.get(url, timeout=TIMEOUT_MAX)
19
20        if response.status_code != 200:
21            print("Error: ", response.status_code)
22            return None
23
24        books_data = response.json()
25        if use_cache:
26            print(f"Saving file {AUTHOR}.json")
27
28        data = response.json()
29        with open(path, "w", encoding="utf-8") as file:
30            json.dump(data, file, indent=2, ensure_ascii=False)

```

```
30     return [Book(**book) for book in books_data]
31
32
33 except requests.exceptions.RequestException as e:
34     print("Error: ", e)
35     return None
```

Podczas tworzenia zbioru stała AUTHOR zawierała napis "adam-mickiewicz". Ten fragment kodu odpowiedzialny jest za pobranie listy książek dla wybranego autora. Book to autorska klasa służąca do deserializacji danych w formacie JSON.

Kod źródłowy 4.2: Fragment kodu odpowiedzialny za pobranie podanej pozycji

```
1 import os
2
3
4 def scrape(epub_url: str, name: str):
5     print("Downloading url: ", epub_url)
6     try:
7         print(f"Getting: {epub_url}")
8         response = requests.get(epub_url, timeout=TIMEOUT_MAX)
9
10        if response.status_code != 200:
11            print("Error: ", response.status_code)
12
13        path = os.path.join(RAW_DIR, f"{name}.epub")
14
15        total_size = int(response.headers.get("content-length", 0))
16        print(f"Downloading {total_size} bytes...")
17
18        with open(path, "wb") as file:
19            if total_size == 0:
20                file.write(response.content)
21            else:
22                downloaded = 0
23                for chunk in response.iter_content(chunk_size=8192):
24                    if chunk:
25                        file.write(chunk)
26                        downloaded += len(chunk)
27                        if total_size > 0:
28                            progress = (downloaded / total_size) *
29                                         100
30                            print(f"Progress: {progress:.1f}%", end=
31                                         "\r")
32
33                print(f"\nSuccessfully saved: {path}")
34
35 except requests.exceptions.RequestException as e:
36     print("Error: ", e)
```

4.2. Ekstrakcja tekstu

Początkowo rozważono wykorzystanie formatów tekstowych, takich jak TXT, HTML oraz XML, które ze względu na swoją prostą, czytelną dla maszyn strukturę, wydawały się obiecującym rozwiązaniem problemu ekstrakcji treści. Proces pozyskania danych z tych formatów sprowadzałby się wówczas do bezpośredniego odczytu i parsowania plików.

Okazało się jednak, że zasoby udostępniane w tych formatach przez serwis WolneLektury.pl są często niekompletne w kontekście całych zbiorów dzieł. Dla zilustrowania tego problemu można posłużyć się przykładem cyklu „Ballady i romanse” Adama Mickiewicza, który składa się z czternastu osobnych utworów. Tymczasem pliki w formatach TXT, HTML i XML dostępne do pobrania dla tego cyklu zawierają wyłącznie tekst pojedynczego wiersza „To lubię”, pomijając pozostałe części dzieła.

Wobec powyższych ograniczeń, ostatecznym wyborem formatu źródłowego do budowy korpusu wybrany został format EPUB. Format ten można traktować jako format binarny, jednakże w rzeczywistości jest to archiwum zawierające ustrukturyzowane pliki (X)HTML, metadane oraz zasoby multimedialne. Taka struktura znaczco komplikuje proces ekstrakcji tekstu w porównaniu z prostymi formatami tekstowymi. Aby uprościć proces wydobywania tekstu z plików EPUB, wykorzystano biblioteki: **EbookLib** do niskopoziomowego parsowania struktury archiwum EPUB oraz **Beautiful Soup** do przetwarzania i ekstrakcji czystego tekstu z sekcji HTML zawartych w książce.

Kod źródłowy 4.3: Fragment autorskiej klasy TextExtractor odpowiedzialnej za ekstrakcję tekstu

```
1 import ebooklib
2 from bs4 import BeautifulSoup
3 from ebooklib import epub
4
5
6 class TextExtractor:
7     def load_file(self, input_path: str):
8         book = epub.read_epub(input_path)
9         self.book_name = input_path.split("/")[-1]
10        items = list(book.get_items_of_type(ebooklib.ITEM_DOCUMENT))
11
12        chapter_items = []
13        for item in items:
14            if "part" in item.get_name().lower():
15                chapter_items.append(item)
16
17        return chapter_items
18
19    def parse_file(self, chapter):
20        soup = BeautifulSoup(chapter.get_body_content(), "html.
21            parser")
22        target_elements = soup.find_all(
23            ["h2", "div"], class_=lambda x: x != "stanza" and x != "
24            stanza-spacer"
25        )
26
27        filtered_elements = []
```

```
26     for element in target_elements:
27         if element.name == "h2":
28             filtered_elements.append(element)
29         elif element.name == "div" and "verse" in element.get("class", []):
30             filtered_elements.append(element)
31
32     text_elements = []
33     for element in filtered_elements:
34         element_text = element.get_text().strip()
35         if element_text:
36             text_elements.append(element_text)
37
38     text = "\n".join(text_elements)
39
40     return text
```

Plik źródłowy zostaje otwarty przy pomocy funkcji z biblioteki `EbookLib`. Otrzymane w ten sposób to pliki składowe archiwum jakim jest format EPUB. Pliki te składają się z wielu elementów, jednakże na potrzeby tej pracy wykorzystano dane z elementów typu "h2" i "div". Zawierają one odpowiednio nagłówki oraz główny tekst z publikacji.

4.3. Przygotowanie obrazów

Obrazy wyjściowe generowane były w rozdzielczości odpowiadającej standardowemu formatowi A4 (2480×3508 pikseli), przy założeniu rozdzielczości 300 DPI (dots per inch). W przyjętej konfiguracji, tekst na obrazach miał rozmiar 60 pikseli a marginesy 100 pikseli. W celu zapewnienia pełnej widoczności tekstu oraz uniknięcia jego przycięcia, podzielono go na fragmenty zawierające x linii tekstu. Wartość x wyliczono za pomocą poniższego wzoru:

$$x = \left\lfloor \frac{H - 2m}{h} \right\rfloor$$

Gdzie:

- H - Wysokość obrazu (3508 px)
- m - Margines (100 px)
- r - Rozmiar tekstu (60 px + odstęp między wierszowy)

Do generacji obrazów wykorzystano bibliotekę `Pillow`, jest to standardowa biblioteka do generacji obrazów w języku Python. Zaletą wykorzystania tej biblioteki jest prostota zmiany kroju pisma użytego do generacji. `Pillow` udostępnia klasę `FreeTypeFont`, która pozwala na załadowanie pliku w formacie ttf. Następnie załadowany krój pisma można przekazać jako argument podczas rysowania tekstu.

Kod źródłowy 4.4: Fragment autorskiej klasy `ImageGenerator` odpowiedzialnej za generację obrazów

```
1 | from PIL import Image, ImageDraw
```

```
2
3
4 class ImageGenerator:
5     def text_to_img(self, output_path, text: str):
6         output_path = output_path.replace(".txt", ".png")
7         lines = text.split("\n")
8
9         img = Image.new("RGB", (self.A4_WIDTH, self.A4_HEIGHT),
10                      color="white")
11         draw = ImageDraw.Draw(img)
12
13         y_position = self.margin
14
15         for line in lines:
16             if y_position + self.line_height < self.A4_HEIGHT - self
17                         .margin:
18                 draw.text((self.margin, y_position), line, fill="
19                               black", font=self.font)
20                 y_position += self.line_height
21
22         img.save(output_path)
```


5. Aspekt badawczy

Pytania badawcze

1. Który algorytm zapewnia najlepszy kompromis między dokładnością a szybkością przetwarzania?
2. W jakim stopniu opracowany autorski zbiór danych jest wykazuje porównywalność z uznanymi benchmarkami OCR pod względem trudności i zdolności do weryfikacji podstawowej skuteczności algorytmów?
3. Czy zastosowane kategorie znieksztalcen wizualnych wpływają w sposób statystycznie istotny na efektywność działania algorytmów OCR?
4. Czy wybrane kroje pisma mają statystycznie istotny wpływ na efektywność działania wybranych algorytmów?

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 Wilcooxona 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).

5.1. Scenariusze eksperymentów

Eksperiment 1 Porównanie szybkości przetwarzania i jakości wyników

Celem pierwszego eksperymentu była analiza związku między czasem przetwarzania a jakością otrzymanych wyników. Eksperiment został przeprowadzony na czystej wersji (bez znieksztalcen oraz krój pisma Ubuntu) autorskiego zbioru danych. Eksperiment przeprowadzono w celu odpowiedzi na pytanie badawcze nr. 1.

Eksperiment 2 Porównanie autorskiego zbioru z uznanymi benchmarkami

Celem tego eksperymentu było 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 została wykorzystana czysta wersja autorskiego zbioru. Eksperiment przeprowadzono w celu odpowiedzi na pytanie badawcze nr. 2. Wyniki zostały porównywane na podstawie współczynnika błędów znakowych (CER).

Eksperiment 3 Wpływ znieksztalcen obrazu

Trzeci eksperiment ma na celu zbadanie wpływu zastosowanych znieksztalcen obrazu (opisane w rozdziale 3) na otrzymane wyniki. Dla każdego rodzaju znieksztalcenia został

wygenerowany nowy fragment autorskiego zbioru danych. Eksperyment zostanie przeprowadzony na tych fragmentach. Eksperyment przeprowadzono w celu odpowiedzi na pytanie badawcze nr. 3. Wyniki zostały porównywane na podstawie metryki współczynnika błędów znakowych (CER).

Eksperyment 4 Wpływ krojów pisma

Natomiast celem czwartego eksperymentu było zbadanie wpływu wybranych krojów pisma (opisane w rozdziale 3) na otrzymane wyniki. Dla każdego kroju pisma został wygenerowany nowy fragment autorskiego zbioru danych. Eksperyment zostanie przeprowadzony na tych fragmentach. Eksperyment przeprowadzono w celu odpowiedzi na pytanie badawcze nr. 4. Wyniki zostały porównywane na podstawie metryki współczynnika błędów znakowych (CER).

6. Wyniki

7. 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, 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] K. Perlin. An image synthesizer. *ACM SIGGRAPH Computer Graphics*, 19(3):287–296, 1985.
- [10] V. S and S. A. Performance comparison of ocr tools. *International Journal of UbiComp*, 6(3):19–30, July 2015.
- [11] 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.