

POLITECHNIKA BIAŁOSTOCKA

WYDZIAŁ INFORMATYKI

PRACA DYPLOMOWA MAGISTERSKA

TEMAT: METODY OCR DO DOKUMENTÓW Z DANYMI TABELARYCZNYMI

WYKONAWCA: PAWEŁ MARCZAK

OPIEKUN PRACY DYPLOMOWEJ : DR INŻ. MIROSŁAW OMIELJANOWICZ

BIAŁYSTOK 2022 ROK

Karta dyplomowa

POLITECHNIKA BIAŁOSTOCKA Wydział Informatyki	Studia stacjonarne stacjonarne/niestacjonarne drugiego stopnia studia pierwszego stopnia/studia drugiego stopnia	Nr albumu studenta 104160
		Rok akademicki 2021/2022
		Kierunek studiów Informatyka Specjalność Biometria i przetwarzanie sygnałów

Paweł Marczak

.....
Imiona i nazwisko studenta

TEMAT PRACY DYPLOMOWEJ:

Metody OCR do dokumentów z danymi tabelarycznymi

Zakres pracy:

1. Przegląd i analiza publikacji
2. Wybór i implementacja kilku metod do przeprowadzenia badań skuteczności działania
3. Zaplanowanie i przeprowadzenie skuteczności działania
4. Podsumowanie badań i wnioski

Słowa kluczowe (max 5): OCR, badanie skuteczności, przetwarzanie obrazów

dr inż. Mirosław Omieljanowicz

.....
Imiona i nazwisko, stopień/tytuł opiekuna pracy dyplomowej - podpis

31.10.2021

30.09.2022

.....
Data wydania tematu pracy dyplomowej
- podpis opiekuna pracy dyplomowej

Regulaminowy termin złożenia pracy dyplomowej

.....
Data złożenia pracy dyplomowej
- potwierdzenie dziekanatu

.....
Ocena opiekuna pracy dyplomowej

.....
Podpis opiekuna pracy dyplomowej

.....
Imiona i nazwisko, stopień/tytuł recenzenta

.....
Ocena recenzenta

.....
Podpis recenzenta

Subject of diploma thesis

Application of OCR methods to documents with tabular data

SUMMARY

The master' s thesis is about applications of computer vision algorithms to the problem of tabular data recognition in digital documents. It was inspired by real – life problem of automating the process of registering invoices in an accounting system. The thesis contains: introduction to theoretical background of the subject, analysis and comparison of approaches popular in literature, author's proposal of table structure recognition (TSR) algorithm based on profile - projection approach, introduction of a modification to a commonly used TSR evaluation method, performance comparison of the developed TSR system with the open-source solution called Camelot and the Nanonets commercial product on a dataset composed of real documents coming from the accounting of a private company.

Paweł Marczak
Imiona i nazwisko studenta
104160
Nr albumu
Informatyka, drugi stopień, stacjonarne
Kierunek i forma studiów
dr inż. Mirosław Omieljanowicz
Opiekun pracy dyplomowej

Białystok, dnia 20.06.2022 r.

OŚWIADCZENIE

Przedkładając w roku akademickim 2021/2022 Opiekunowi pracy dyplomowej dr inż. Mirosławowi Omieljanowiczowi pracę dyplomową pt.: Metody OCR do dokumentów z danymi tabelarycznymi, dalej zwaną pracą dyplomową, **oswiadczam, że:**

- 1) praca dyplomowa stanowi wynik samodzielnej pracy twórczej;
- 2) wykorzystując w pracy dyplomowej materiały źródłowe, w tym w szczególności: monografie, artykuły naukowe, zestawienia zawierające wyniki badań (opublikowane, jak i nieopublikowane), materiały ze stron internetowych, w przypisach wskazywałem/am ich autora, tytuł, miejsce i rok publikacji oraz stronę, z której pochodzą powoływane fragmenty, ponadto w pracy dyplomowej zamieściłem/am bibliografię;
- 3) praca dyplomowa nie zawiera żadnych danych, informacji i materiałów, których publikacja nie jest prawnie dozwolona;
- 4) praca dyplomowa dotychczas nie stanowiła podstawy nadania tytułu zawodowego, stopnia naukowego, tytułu naukowego oraz uzyskania innych kwalifikacji;
- 5) treść pracy dyplomowej przekazanej do dziekanatu Wydziału Informatyki jest jednakowa w wersji drukowanej oraz w formie elektronicznej;
- 6) jestem świadomy/a, że naruszenie praw autorskich podlega odpowiedzialności na podstawie przepisów ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych (Dz. U. z 2021 r. poz. 1062), jednocześnie na podstawie przepisów ustawy z dnia 20 lipca 2018 roku Prawo o szkolnictwie wyższym i nauce (Dz. U. 2021 poz. 478 z późn. zm) stanowi przesłankę wszczęcia postępowania dyscyplinarnego oraz stwierdzenia nieważności postępowania w sprawie nadania tytułu zawodowego;
- 7) udzielam Politechnice Białostockiej nieodpłatnej, nieograniczonej terytorialnie i czasowo licencji wyłącznej na umieszczenie i przechowywanie elektronicznej wersji pracy dyplomowej w zbiorach systemu Archiwum Prac Dyplomowych Politechniki Białostockiej oraz jej zwielokrotniania i udostępniania w formie elektronicznej w zakresie koniecznym do weryfikacji autorstwa tej pracy i ochrony przed przywłaszczeniem jej autorstwa;
- 8) udzielam Politechnice Białostockiej zgódę na wprowadzenie pracy dyplomowej do ogólnopolskiego repozytorium pisemnych prac dyplomowych;
- 9) wyrażam zgodę na umieszczenie mojego imienia i nazwiska, jako autora pracy dyplomowej w systemie ALEPH i OMEGA-Psir (Baza Wiedzy PB) w celu upowszechniania informacji o wypromowanych pracach dyplomowych przez pracowników Uczelni.

.....
czytelny podpis studenta

Spis treści

1 Wstęp	5
2 Wprowadzenie	7
2.1 Zarys historyczny	7
2.2 Podstawy teoretyczne	8
2.3 Automatyczne rozpoznawanie tabeli	11
2.4 Detekcja tabeli	14
2.5 Rozpoznawanie struktury tabeli	15
2.5.1 Ewaluacja	17
2.5.2 Systemy TSR wybrane z literatury	19
3 Przeprowadzone eksperymenty i badania	22
3.1 Usczegółowienie wymagań i przyjęte założenia	22
3.2 Przygotowana baza dokumentów tabelarycznych	24
3.3 Opracowana metoda rozpoznawania struktury tabel	30
3.3.1 Wybór podejścia	30
3.3.2 Własne modyfikacje algorytmów	41
3.3.3 Opracowane narzędzie	42
3.3.4 Zarys zaproponowanej metody rozpoznawania struktury	44
3.3.5 Detekcja obramowań	46
3.3.6 Wnioskowanie bez korzystania z obramowań	52
3.3.7 Łączenie wyników	70
3.4 Porównywane metody ekstrakcji tabel	71
3.5 Wybrane metody ewaluacji	73
3.6 Zestawienie i analiza wyników	77
4 Podsumowanie	86
4.1 Spostrzeżenia i wnioski	86
4.2 Przyszłe prace	89

1. Wstęp

Współczesność oparta jest na przetwarzaniu danych. Trudno jest sobie obecnie wyobrazić funkcjonowanie wielu gałęzi nauki i przemysłu bez dostępu do nowoczesnych systemów informatycznych usprawniających przepływ informacji. W istocie, praca informatyka oparta jest o przetwarzanie informacji w różnych formach i stopniach ustrukturyzowania, zwłaszcza dziś, przy rosnącym zapotrzebowaniu na produkty wykorzystujące sztuczną inteligencję i uczenie maszynowe. Wraz z rozwojem technologii informatycznych, systemy są przystosowywane do coraz skuteczniejszego pozyskiwania danych z różnych źródeł i przystosowywania ich do użytku zgodnie z zapotrzebowaniem.

Szczególnym przykładem źródła informacji są cyfrowe wersje dokumentów (zeskanowanych lub wygenerowanych cyfrowo egzemplarzy). W artykułach naukowych, książkach, dokumentach księgowych i raportach często znajduje się wiedza, której wykorzystanie może przynieść wymierne korzyści. Kluczowe w ewentualnej automatyzacji ekstrakcji takich danych jest zachowanie relacyjnej struktury danych, często zakodowanej w postaci tabeli.

Celem niniejszej pracy jest udowodnienie tezy, że możliwe jest pozyskanie danych tabelarycznych z dokumentów elektronicznych w formie plików binarnych z odpowiednio dużą skutecznością (przekraczającą 90% dokładności na zaproponowanym zbiorze testowym) uzasadniającą praktyczne wykorzystanie rezultatów. Realizacja w/w celu zostanie osiągnięta poprzez:

- analizę krytyczną publikacji dotyczących przetwarzania dokumentów zawierających dane tabelaryczne oraz pozyskanie zestawu dokumentów do wykonania eksperymentów,
- opracowanie i implementację wybranych algorytmów rozpoznawania struktury tabeli,
- wprowadzenie własnych modyfikacji wynikających z przeprowadzonych analiz,
- przeprowadzenie badań skuteczności przygotowanych rozwiązań w zestawieniu z wybranymi systemami analizującymi dane tabelaryczne.

Istotnym faktem jest to, że temat pracy powstał z inicjatywy jej autora, a uzyskane efekty docelowo mają posłużyć w usprawnieniu procesów księgowych w prywatnym przedsiębiorstwie.

Niektóre decyzje dotyczące zakresu prac były podejmowane (w porozumieniu z promotorem) w celu wsparcia ewentualnego wdrożenia opracowanego rozwiązania w konkretnej firmie.

Należy także stwierdzić, że takie podejście do problemu (nakierowanie prac na praktyczne użycie) nie jest czymś niespotykanym w literaturze i jest sensowne, zważając na złożoność i praktyczność zagadnienia [4], [34].

2. Wprowadzenie

2.1 Zarys historyczny

Tabela, metoda wizualnego porządkowania informacji, była wykorzystywana już przez starożytnych Sumerów ok. 4500 lat temu, służąc im do utrwalania danych matematycznych na kamiennych tabliczkach. Od tego czasu technologie uległy dużym zmianom, ale sam koncepcja prezentowania ustrukturyzowanych danych w uporządkowany graficznie sposób służy człowiekowi aż do dziś [13].

Sama graficzna reprezentacja danych w praktyce czasami okazywała się niewystarczająca. Dane tabelaryczne wymagały dodatkowego przetwarzania i interpretacji, by mogły być zastosowane przy zaawansowanych analizach i prognozach. Ręczne wykonanie tych operacji jest uciążliwe i czasochłonne. Wraz z rozwojem technik informatycznych pojawiły się koncepcje maszynowej ekstrakcji danych z tabel.

Problem automatycznej ekstrakcji danych tabelarycznych z dokumentów był zaadresowany już w latach 80 - tych jako podproblem rozpoznawania struktury dokumentu w zyskującej na popularności dziedzinie przetwarzania dokumentów cyfrowych. Było to zagadnienie niezwykle ważne, ze względu ogromne zapotrzebowanie na tego typu rozwiązania w różnych branżach działalności gospodarczych. Dużą przeszkodą, stojącą w tamtym czasie na drodze postępu badawczego, było niejednoznaczne zdefiniowanie problemu ze względu na różnorodność nośników danych [3], [34].

Aż do końca XX wieku nie było jednego ustalonego formatu przechowywania dokumentów cyfrowych, w zależności od instytucji korzystano z różnych rozwiązań. Niestety, większość z nich borykała się z problemami przy wyświetlaniu na różnych platformach. Z tego powodu, w starszych pracach badawczych rzadko korzystano z dedykowanych dla dokumentów formatów, w zamian pracowano ze zdjęciami wspartymi technologią OCR (która w tamtych czasach często była zawodna) lub przedstawiano jedynie teoretyczną koncepcję rozwiązania, które później mogło być wdrożone przez osoby zainteresowane w prywatnych systemach [17], [21], [22]. Należy zaznaczyć, że z czasem techniki widzenia komputerowego rozwinęły się do tego stopnia, że osiągane przez nie wyniki nie odstawały mocno od wyników systemów wykorzystujących zakodowane w plikach dokumentów dodatkowe informacje, zwane metadanymi [25], [29], [27], [20], [9].

W roku 2008 format PDF stał się standardowym formatem przechowywania dokumentów cyfrowych przyjętym przez ISO. Wszystkie liczące się edytory tekstu posiadały opcję eksportu do tego formatu z zachowaniem dokładnej formy graficznej i pełnej zawartości tekstowej. Został on zdecydowanie najczęściej stosowanym formatem, wykorzystywany od biur po szkoły i instytucje, niezmiennie liderując do dnia dzisiejszego. Bez wątpienia jego największą zaletą była przenosalność - był wyświetlany jednakowo niezależnie od urządzenia i systemu operacyjnego [31].

Analizując artykuły przeglądowe [34],[5],[18], wyraźnie można wyróżnić kilka trendów funkcjonujących w dziedzinie automatycznego rozpoznawania tabel na przestrzeni ostatnich 40 lat. Początkowo (lata 1980-2000) systemy skupiały się głównie na ekstrakcji cech geometrycznych z obrazów dokumentów (inaczej metody top-down) [20], [17], [1], [36]. Były w większości oparte na mniej lub bardziej złożonych algorytmach regułowych, często dostosowanych do konkretnych typów tabel. W latach 2000 - 2016, wraz z rozwojem technik OCR i wejściem do kanonu formatu PDF, coraz częściej wykorzystywane były metody opierające się na detekcji słów na stronie (inaczej bottom-up), analizujące najbardziej uniwersalną z cech dla zróżnicowanych stylów tabel, czyli układ tekstu [22], [35], [9]. W ostatnich latach ogromną popularność zyskało podejście skoncentrowane na rozpoznawaniu złożonych struktur z wykorzystaniem głębokiego uczenia i sieci CNN [25], [29], [27] , [12], [19]. Zmieniające się trendy nie wpływają na fakt, że każdy z nich ma swoje mocne i słabe strony oraz, że starsze podejścia znajdują swoje zastosowania również w publikacjach z ostatnich lat [2], [20], [27], [35].

2.2 Podstawy teoretyczne

Pojęcie tabeli jest powszechnie znane, większość osób posługuje się tabelami w sposób intuicyjny. Posiłkując się informacjami przytoczonymi przez Coüasnona [5] i Zanibbiego [34], sformułowano ogólny opis pojęcia tabeli, który dobrze oddawał charakter przypadków, na które natknieto się podczas realizacji pracy.

Generalnie, tabelę można nazwać rozpowszechnionym środkiem wizualnej reprezentacji ustrukturyzowanych danych. Reprezentacja ta może zawierać słowa, liczby, wzory, grafiki, a nawet zagnieżdżone tabele. Elementy te mogą być wydrukowane lub spisane ręcznie. Składa się z komórek, których typ danych zależy od ich pionowego i poziomego indeksu.

Bardziej formalnie - tabela jest kratową reprezentacją macierzy M_{mn} , gdzie każdy element macierzy m_{ij} ($i \leq m, j \leq n$) jest atomowy (nie dający się podzielić). Taki atomowy podział tabeli często nazywany jest kratą lub siatką tabeli (ang. *grid*).

Należy zaznaczyć, że koncepcja tabeli zgodna z tą definicją nie zawsze musi mieć dokładne przełożenie w jej wizualnej reprezentacji, którą często cechują różnego typu nieregularności, takie

jak zespalone lub zagnieżdżone komórki lub puste obszary. Te elementy ogólnej struktury tabeli w odniesieniu do jej kraty opisywane są jako układ tabeli (ang. *layout*).

W układzie tabeli można wyróżnić części składowe pojawiające się wtórnie w różnych tabelach i pełniące tę samą, określona funkcję. Elementy te nazywa się regionami tabeli.

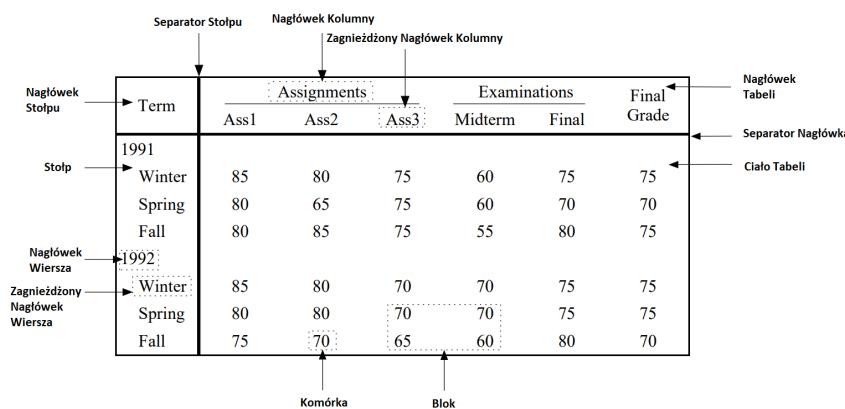
Z definicji, przy wyodrębnianiu kraty, układu i regionów tabeli muszą być pomocne optyczne przesłanki (np. obramowania, wolne przestrzenie, wyrównanie tekstu, kolor tła).

Journal	Full Name	Details	
		Appears	Publisher
TPAMI	IEEE Transactions on Pattern Analysis and Machine Intelligence	monthly	IEEE
IJDAR	International Journal on Document Analysis and Recognition	quarterly	Springer-Verlag
PR	Pattern Recognition	monthly	Elsevier
IJPRAI	International Journal on Pattern Recognition and Artificial Intelligence	eight times/year	World Scientific

Rysunek 2.1: Przykład tabeli o złożonym układzie, rysunek pochodzący z publikacji Zanibbiego [34]

A1–A2	B1–B2	C1–D1	
		C2	D2
A3	B3	C3	D3
A4	B4	C4	D4
A5	B5	C5	D5
A6	B6	C6	D6

Rysunek 2.2: Opis układu tabeli z rysunku 2.1 w odniesieniu do jej kraty (wiersze kraty są indeksowane od 1 do 6, kolumny od A do D), rysunek pochodzący z publikacji Zanibbiego [34]



Rysunek 2.3: Wyszczególnienie regionów tabeli, rysunek pochodzący z publikacji Zanibbiego [34] i zmodyfikowany przez autora

Dwie tabele mogą cechować się taką samą kratą, układem, regionami i zawartością, a mimo to graficznie różnić się od siebie. Wszystkie czynniki, z których wynikają te różnice można określić szerokim pojęciem stylu tabeli. Przykładami takich czynników mogą być, np. styl czcionki zawartości tekstuowej, typy separatorów komórkowych czy sposób wyrównywania tekstu.

Dwuwymiarowe tabele często dzieli się na dwa rodzaje - złożone i proste [5]. Tabele proste mają wyraźnie wyeksponowaną strukturę kraty, każdy zestaw indeksów w płaszczyźnie poziomej i pionowej wskazuje jednoznacznie na konkretną i niepowtarzalną komórkę. Dla uproszczenia, można przyjmować, że ich układ jest równoznaczny ich kracie (rys. 2.4).

Tabele złożone najprościej określić jako tabele nieproste - ich struktura jest mniej regularna, występują zespółone bądź puste komórki, dodatkowe zagnieżdżone poziomy w nagłówkach utrudniają identyfikację ścieżki dostępu do konkretnych komórek, a także pojawiają się inne problemy utrudniające dekompozycję struktury (rys. 2.5).

Lp	Symbol	Nazwa	Kod kreskowy	Ilość	j.m.	Rabat [%]	Cena netto	VAT [%]	Wartość netto	VAT	Wartość brutto
20	CTA4/NIEB	TRES Clipboard teczka A4 niebieski	5901878231082	1,000	szt.	30,00	6,10	23	6,10	1,40	7,50
21	CTA4/SZAR	TRES Clipboard teczka A4 szary	5901878231624	1,000	szt.	30,00	6,10	23	6,10	1,40	7,50
22	CTA4/ZOL	TRES Clipboard teczka A4 złoty	5901878231549	1,000	szt.	30,00	6,10	23	6,10	1,40	7,50
23	DRBI/120	TRES Nici Iniane dratwa bielona 120 MB 10 dag	5901878232140	8,000	szt.	30,00	7,95	23	63,60	14,63	78,23
24	SKNA4/CZAR	TRES skonoszty sztywny niewpiniany A4 PCV czarny	5901878231365	80,000	szt.	30,00	0,60	23	48,00	11,04	59,04
25	SKNA4/CZE	TRES skonoszty sztywny niewpiniany A4 PCV czerwony	5901878231327	50,000	szt.	30,00	0,60	23	30,00	6,90	36,90
26	SKNA4/GRAN	TRES skonoszty sztywny niewpiniany A4 PCV granatowy	5901878231402	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
27	SKNA4/SZARY	TRES skonoszty sztywny niewpiniany A4 PCV szary	5901878231396	10,000	szt.	30,00	0,60	23	6,00	1,38	7,38
28	SKNA4/ZIE	TRES skonoszty sztywny niewpiniany A4 PCV zielony	5901878231334	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
29	SKNA4/BŁĘ	TRES skonoszty sztywny niewpiniany A4 PCV błękitny	5901878238838	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
30	SKWA4/CZAR	TRES skonoszty sztywny wpinany A4 PCV czarny	5901878231426	80,000	szt.	30,00	0,60	23	48,00	11,04	59,04
31	SKWA4/FIO	TRES skonoszty sztywny wpinany A4 PCV fioletowy	5901878238296	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
32	SKWA4/CZER	TRES skonoszty sztywny wpinany A4 PCV czerwony	5901878231433	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
33	SKWA4/GRAN	TRES skonoszty sztywny wpinany A4 PCV granat	5901878231495	30,000	szt.	30,00	0,60	23	18,00	4,14	22,14
34	SKWA4/LIM	TRES skonoszty sztywny wpinany A4 PCV limonkowy	5901878238890	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
35	SKWA4/NIE	TRES skonoszty sztywny wpinany A4 PCV niebieski	5901878231440	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
36	SKWA4/POM	TRES skonoszty sztywny wpinany A4 PCV pomarańcz	5901878231471	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
37	SKWA4/RÓŻ	TRES skonoszty sztywny wpinany A4 PCV różowy	5901878238289	20,000	szt.	30,00	0,60	23	12,00	2,76	14,76
38	SKWA4/SZAR	TRES skonoszty sztywny wpinany A4 PCV szary	5901878231488	40,000	szt.	30,00	0,60	23	24,00	5,52	29,52
39	SKWA4/ZOL	TRES skonoszty sztywny wpinany A4 PCV złoty	5901878231464	40,000	szt.	30,00	0,60	23	24,00	5,52	29,52
40	SZOZMIX/10DAG	TRES Sznurek ozdobny 10 dag 50 mb mix kolor AG	5901878238302	5,000	szt.	30,00	4,07	23	20,35	4,68	25,03
41	SZJU/10DAG	TRES Sznurek szpagat jutowy 40 MB 10 dag	5901878234793	6,000	szt.	30,00	2,68	23	16,08	3,70	19,78

Rysunek 2.4: Przykład tabeli prostej pochodzącej z autorskiego zbioru danych

Lp.	Symbol Nazwa produktu	Ilość	J.m.	Rabat %	Cena netto	Kwota Netto	Podatek VAT %	Kwota Brutto
17	TC96/1 MB TABLICA KORKOWA 90X60CM W RAMIE DREWNIANEJ MEMOBOARDS [TC96/1 MB]	5903273024658	2 szt	25,00	15,26	30,52 P23	7,02	37,54
18	TC75/1 MB TABLICA KORKOWA 70X50CM W RAMIE DREWNIANEJ [TC75/1 MB]	5903273028113	1 szt	25,00	15,50	15,50 P23	3,57	19,07
19	TC85/1 MB TABLICA KORKOWA 80X50CM W RAMIE DREWNIANEJ [TC85/1 MB]	5903273032523	2 szt	25,00	16,34	32,68 P23	7,52	40,20
20	TM64ALC/1 MB TABLICA SUCHÓŚCIERALNO-MAGNETYCZNA 60X40CM W RAMIE ALUMINIOWEJ CLASSI [TM64ALC/1 MB]	5903273027161	1 szt	25,00	30,13	30,13 P23	6,93	37,06
21	TS96/1 MB TABLICA SUCHÓŚCIERALNA 90X60CM W RAMIE DREWNIANEJ [TS96/1 MB]	5903273037207	1 szt	25,00	21,12	21,12 P23	4,86	25,98
22	TS34/1 MB TABLICA SUCHÓŚCIERALNA 30X40CM W RAMIE DREWNIANEJ [TS34/1 MB]	5903273039676	1 szt	25,00	8,29	8,29 P23	1,91	10,20
23	154821 FC ZAKREŚLACZ 48 CZERWONY FABER-CASTELL [154821 FC]	4005401548218	10 szt	20,00	1,99	19,90 P23	4,58	24,48
24	154851 FC ZAKREŚLACZ 48 NIEBIESKI FABER-CASTELL [154851 FC]	4005401548515	10 szt	20,00	1,99	19,90 P23	4,58	24,48
25	154815 FC ZAKREŚLACZ 48 POMARAŃCZOWY FABER-CASTELL [154815 FC]	4005401548157	20 szt	20,00	1,99	39,80 P23	9,15	48,95
26	154828 FC ZAKREŚLACZ 48 RÓŻOWY FABER-CASTELL [154828 FC]	4005401548287	10 szt	20,00	1,99	19,90 P23	4,58	24,48
27	154863 FC ZAKREŚLACZ 48 ZIELONY FABER-CASTELL [154863 FC]	4005401548638	20 szt	20,00	1,99	39,80 P23	9,15	48,95
28	154807 FC ZAKREŚLACZ 48 ZŁOTY FABER-CASTELL [154807 FC]	4005401548072	20 szt	20,00	1,99	39,80 P23	9,15	48,95
29	7/5-BL ED TEXTMARKERY EDDING MINI 5 KOL. BLISTERER [7/5-BL ED]	4004764958795	1 kpl	20,00	6,39	6,39 P23	1,47	7,86
30	020-1760 NO ZSZYWACZ NOVUS E15 CZARNY Z PAKIETEM STARTOWYM ZSZYWEK No.10 DIN SUPER [020-1760 NO]	4009729048597	1 szt	20,00	7,48	7,48 P23	1,72	9,20
31	020-1474 NO ZSZYWACZ NOVUS C2 NIEBIESKI Z PAKIETEM STARTOWYM ZSZYWEK 24/6 DIN SUPER [020-1474 NO]	4009729020821	1 szt	20,00	18,39	18,39 P23	4,23	22,62
32	020-1472 NO ZSZYWACZ NOVUS C2 CZARNY Z PAKIETEM STARTOWYM ZSZYWEK 24/6 DIN SUPER [020-1472 NO]	4009729020876	4 szt	20,00	18,39	73,56 P23	16,92	90,48
33	020-1725 NO ZSZYWACZ NOVUS STABIL NIEBIESKI [020-1725 NO]	4009729046951	2 szt	20,00	12,47	24,94 P23	5,74	30,68
34	020-1287 NO ZSZYWACZ NOVUS STABIL CZARNY [020-1287 NO]	4009729009482	2 szt	20,00	12,47	24,94 P23	5,74	30,68

Rysunek 2.5: Przykład tabeli złożonej pochodzącej z autorskiego zbioru danych

Przedstawiona charakterystyka tabel nie wchodzi szczegółowo w niuanse dotyczące zagadnienia i pozostawia pewne pole do interpretacji. Podchodząc bardziej rygorystycznie do badania struktur tabelarycznych, wartą przeczytania jest rozprawa doktorska Wanga [33]. Jej autor w sposób uporządkowany rozwija formalne definicje dotyczące tabel i wyprowadza ich poszczególne właściwości w oparciu o skonstruowane modele matematyczne. Tak formalne podejście nie jest jednak konieczne z punktu widzenia niniejszej pracy magisterskiej, ze względu na jej charakter praktyczny.

2.3 Automatyczne rozpoznawanie tabeli

Automatyczne rozpoznawanie tabeli (lub po prostu rozpoznawanie tabeli) jest zagadnieniem ściśle powiązanym z inteligentnym przetwarzaniem dokumentów, dziedziną informatyki zajmującą się zautomatyzowanym pozyskiwaniem i przetwarzaniem danych zawartych w cyfrowych wersjach dokumentów różnego typu. Zazwyczaj, celem systemów rozpoznających tabele jest wyszukanie i przekonwertowanie tabel z cyfrowej wersji dokumentu lub innych podobnych źródeł do postaci łatwo interpretowanej przez specjalistyczny software. Takimi postaciami mogą być przykładowo

pliki XML, HTML, XLSX czy CSV.

Zdarza się, że systemy analizujące tabele mają również bardziej zaawansowane zadania, takie jak identyfikacja stylu tabeli, czy klasyfikacja jej regionów [11]. Zakres zadań algorytmu może zostać dopasowany pod konkretną aplikację.

W literaturze często rozróżnia się metody rozpoznawania tabeli na dwa rodzaje - wykorzystujące metadane zawarte w plikach źródłowych (przeważnie PDF) i oparte jedynie na analizie końcowej reprezentacji graficznej (obrazu) [18], [32]. Te drugie są przeważnie uważane za bardziej skomplikowane, ze względu na ograniczony dostęp do danych, którymi się posługują (podczas gdy korzystanie z metadanych nie ogranicza korzystania również z cech graficznych z obrazu). Naturalnie, w praktyce dokumenty nie zawsze muszą zapewniać dostęp do odpowiednich metadanych dopasowanych do założeń systemu, przez co operowanie na obrazach jest bardziej uniwersalne.

Od strony technicznej, problem rozpoznawania tabeli rozkłada się na dwa główne podproblemy - detekcję tabeli i rozpoznanie jej struktury. Oba podproblemy są skomplikowane, przez co wiele publikacji skupia się tylko na jednym z nich [10], [20], [4], [12], [22], [35], [36]. Nie jest to w pełni uzasadnione podejście, gdyż dla osiągnięcia przydatnych rezultatów systemy te powinny ściśle współpracować (detekcja tabeli bez rozpoznania jej struktury rzadko kiedy jest pożyteczna, z kolei skuteczność rozpoznawania struktury jest ściśle uzależniona od wyników detekcji). Możliwe jest jednak zaprojektowanie systemu, tak by zajmował się tylko jednym z zagadnień, zostawiając drugie manualnej obróbce użytkownika [2].

W literaturze wyróżnia się trzy ogólne podejścia do problemu rozpoznawania tabeli, zarówno kroku detekcji, jak i rozpoznawania struktury [5], [18], [35]. Są to podejścia bottom - up, top - down i głębokie uczenie z wykorzystaniem CNN.

Podejścia top - down [20], [17], [1], [36] korzystają w głównej mierze z manualnej ekstrakcji cech geometrycznych z obrazu, takich jak detekcja obramowań, analiza marginesów czy zagęszczania czarnych pikseli w regionach obrazu. Metody oparte na tym podejściu stosunkowo łatwo jest dostroić do wyszukiwania konkretnych optycznych przesłanek zawartych w rozpatrywanych tabelach, ale z uwagi na dużą różnorodność stylów wśród ogółu przypadków z życia codziennego, często mają one problemy uniwersalnym, stabilnie dobrym działaniem. Charakterystycznym dla tego podejścia jest to, że ekstrakcja słów zawartych w tabeli jest wykonywana po zakończeniu głównego procesu wnioskowania (może następować po niej ewentualny post - processing).

W metodach opartych na podejściu bottom - up [22], [35], [9], w kontraste do top - down, ekstrakcja słów jest wykonywana na początku procesu rozpoznawania. Wykryty tekst jest swego rodzaju silnikiem metody, której zadaniem jest jego klasteryzacja i wywnioskowanie relacji sąsiedztwa pomiędzy znalezionymi grupami. Jak nie trudno się domyślić, moduł ekstrakcji słów jest w tym wypadku kluczowy dla ogólnej jakości działania systemu, przez co podejście to jest szczególnie popularne wśród rozwiązań korzystających z metadanych.

	Top - down	Bottom - up	DL CNN
Zalety	<ul style="list-style-type: none"> dobre wyniki dla tabel o nieskomplikowanej strukturze skuteczne przy konkretnym stylu/modelu tabeli 	<ul style="list-style-type: none"> skuteczne również dla złożonych układów tabeli możliwość wykorzystania danych semantycznych przy wnioskowaniu struktury stosunkowo uniwersalne 	<ul style="list-style-type: none"> skuteczne nawet dla bardzo skomplikowanych struktur tabel dopasowane do potrzeb rynku komercyjnego
Wady	<ul style="list-style-type: none"> mała uniwersalność skomplikowanie modułów decyzyjnych dla bardziej złożonych przypadków 	<ul style="list-style-type: none"> silnie uzależnione od jakości i ilości przygotowanych danych treningowych natywnie nie wykorzystują dodatkowych separatorów graficznych, np. linii lub kolorów tła 	<ul style="list-style-type: none"> zróżnicowanie metod, problemy interpretacyjne i pracochłonność powodują wiele problemów w sporządzaniu wartościowego ground - truth

Tablica 2.1: Zestawienie typów metod rozpoznawania tabel sporządzone przez autora na podstawie wniosków z zebranej literatury

Ostatnim podejściem, najnowszym z wymienionych, jest analiza tabel z wykorzystaniem głębokich sieci CNN [25], [25], [27], [12], [19]. Metody oparte na tych sieciach często równolegle podchodzą do problemów detekcji i rozpoznawania struktury, przez co nazywa się je rozwiązaniami end - to - end. Sieci CNN samodzielnie wybierają sposób, w jaki podchodzą do ekstrakcji cech z obrazu, przez co nie da się ich jednoznacznie zaliczyć do jednej z dwóch powyższych kategorii. Jest to podejście najprężniej rozwijające się w ostatnich latach, ze względu na popularność głębokiego uczenia w przemyśle, obiecujące rezultaty osiągane przez reprezentujące je publikacje i rosnące zapotrzebowanie na rozwiązania oparte na danych [18]. Duże uzależnienie od ilości i jakości danych jest niestety obecnie również ograniczeniem podejścia, ze względu na małą liczbę i stosunkowo niewielkie objętości publicznych zbiorów danych z odpowiednio przygotowanym ground - truth (fakt, że konkretne rozwiązania mogą wymagać dedykowanej formy adnotacji jeszcze bardziej komplikuje zagadnienie). Z tego powodu testowane było wykorzystanie transfer - learningu, czyli importowanie podstaw sieci z wstępnie wytrenowanymi na innym problemie wagami, następnie modyfikowanie jej i dotrenowywanie na dokumentach zawierających tabele [25], [29], [27].

Trzy wyżej wymienione podejścia są najczęstszym sposobem klasyfikacji konkretnych me-

tod. W praktyce, algorytmy często wykorzystują podejścia hybrydowe, by lepiej radzić sobie z słabymi stronami poszczególnych ujęć tematu. Przykładami takiej hybrydyzacji, może być np. wykorzystanie różnych podejść w różnych trybach działania systemu [2] lub równolegle korzystanie przy wnioskowaniu z cech wydobytych na różne sposoby [17], [25].

2.4 Detekcja tabeli

Celem detekcji tabeli jest zwyczajowo wyznaczenie koordynatów prostokąta zawierającego tabelę w układzie współrzędnych odpowiadających reprezentacji graficznej strony dokumentu. Tak sformułowany problem ma wiele wspólnego z klasycznym problemem detekcji obiektów na obrazie i wiele rozwiązań faktycznie tak do niego podchodzi.

Wykorzystywane jest wiele różnych podejść do problemu, poczynając od klasyfikacji linii tekstu pozyskanych z wykorzystaniem metadanych [10], poprzez analizę połączonych komponentów z wykorzystaniem ręcznie wydobytych cech geometrycznych i technik uczenia maszynowego nadzorowanego [20], po metody głębokie, bardzo często wykorzystywane obecnie do zadań detekcji obiektów na obrazie [25], [29], [29], [19].

Często wykorzystywane modele głębokie to Cascade Mask R-CNN [27], Faster R-CNN [29], YOLO [19] i F - CNN klasyfikujące piksele na obrazie [25]. Podkreślenia wymaga fakt, że metody oparte na konwolucyjnych sieciach neuronowych, według przeglądu z 2021 roku [18], osiągają bardzo dobre wyniki, predysponujące je do praktycznego wykorzystania.

Metoda	Zbiór danych	F - measure	Podejście
DeepDeSRT [29]	ICDAR-2013	96.7	Faster R - CNN
Huynh [20]	ICDAR-2013	98.15	CCA + SVM & Random Forest
CascadeTabNet [27]	ICDAR-2013	100	Cascade Mask R - CNN
Huang [19]	ICDAR-2013	97.3	YOLO
TableNet [25]	ICDAR-2013	96.6	F - CNN

Tablica 2.2: Zestawienie wybranych metod z literatury pod kątem wyników detekcji tabel

Generalnie, analizując artykuł można wysnuć wniosek, że podproblem detekcji w rozpoznawaniu tabeli jest tym łatwiejszym i został już w dużym stopniu rozwiązany. Same metodologie rozwiązań oraz ich ewaluacji zostały też w znacząco ustandaryzowane, przez co temat jest mniej ciekawy oraz angażujący badawczo od podproblemu rozpoznawania struktury. Dodatkowo, manualne zaznaczenie obszaru tabeli jest zdecydowanie mniej czasochłonne od ręcznego wyznaczenia segmentacji strukturalnej. Z wymienionych powodów, autor postanowił w większym stopniu skupić się na tym drugim zagadnieniu w ramach niniejszej pracy magisterskiej.

2.5 Rozpoznawanie struktury tabeli

Rozpoznawanie struktury tabeli (TSR) jest problemem bardziej niejednoznaczny i technicznie skomplikowanym, przez co stanowi większe wyzwanie. Już to w jakim zakresie ma docelowo działać metoda i jaki ma być wynik jej przetwarzania ma duży wpływ na wykorzystane rozwiązania. Dokładając do tego dużą różnorodność stylów i układów tabel, przyjęcie pewnych założeń od początku procedury przetwarzania wydaje się być nieuniknione.

Przyjmowanie określonych założeń jest zabiegiem znany i powszechnie stosowanym w literaturze. Rozwiążanie Clinchanta [4] odczytuje pozycje kolumn z ground - truth. Biblioteka Camelot [2] wykorzystuje dodatkowe parametry dostrajające podane przez użytkownika. DeepDESRT [29] oraz rozwiązanie komercyjne Nanonets przyjmują, że rozpoznanie struktury sprowadza się do wyznaczenia jej siatki. Na podobnej zasadzie działa rozwiązanie TableNet [25], przy czym posiada mocno ograniczony moduł segmentacji wierszowej.

W przeglądzie Zanibbiego [34] zestaw przyjętych założeń został określony modelem rozpoznawanej tabeli. Obranie konkretnego modelu poprawia wyniki, pozwala na uporządkowaną ewaluację zaproponowanej metody i późniejsze praktyczne jej zastosowanie dla konkretnego problemu. Odbywa się to kosztem uniwersalności - zależnie od odgórnych ustaleń nie każdy typ tabeli musi być poprawnie interpretowany lub wynik działania algorytmu może mieć zawężone pole zastosowania.

Najnowsze rozwiązania startujące w konkursach przyjmują stosunkowo ogólny model tabeli, uwzględniający zarówno siatkę jak i układ tabel. Taki model wymaga segmentacji na poziomie komórkowej i dodatkowego modułu wnioskującego relacje ich sąsiedztwa. Wynikiem dokonanego w ten sposób przetwarzania zwyczajowo jest plik XML o określonej strukturze. W konkursie ICDAR - 2013 wymagane było uwzględnienie w rezultacie pełnej informacji o układzie, zawartości i koordynatach komórek tabel [15]. Dla późniejszej edycji tego konkursu z 2019 roku, skierowanej do rozwiązań opartych jedynie na analizie obrazu, informacje o zawartości tekstuowej nie były wymagane, by nie penalizować nadmiernie błędów, spowodowanych działaniem modułu OCR [8].

Należy zaznaczyć, że forma wyniku wymagana w 2013 roku może być praktycznie wykorzystywana, przykładowo poprzez przekonwertowanie go do postaci łatwego w parsowaniu pliku CSV (np. zgodnie z założeniami wykorzystywany przez bibliotekę Camelot [2] lub bezpośrednie przeparsowanie go w systemie specjalistycznym. Z drugiej strony, forma wymagana w 2019 roku nie zawiera żadnych informacji o zawartości tabel, przez co sama w sobie, bez dokumentu wejściowego i dodatkowego modułu OCR nie ma zastosowania.

```
<?xml version="1.0" encoding="UTF-8"?>
<document filename='filename.pdf'>
  <table id='0'>
    <region id='0' page='3' col-increment='0' row-increment='0'>
      <cell id='0' start-row='0' start-col='0'>
        <bounding-box x1='70' y1='79' x2='131' y2='91'>
        <content>COUNTRY</content>
        <instruction instr-id='65' subinstr-id='0'>
        </cell>
      <cell id='1' start-row='0' start-col='1' end-col='2'>
        <bounding-box x1='165' y1='79' x2='201' y2='91'>
        <content>3 years</content>
        <instruction instr-id='65' subinstr-id='2'>
        </cell>
      <cell id='2' start-row='0' start-col='3'>
        <bounding-box x1='234' y1='79' x2='271' y2='91'>
        <content>4 years</content>
        <instruction instr-id='65' subinstr-id='4'>
        </cell>
      ...
    </region>
    ...
  </table>
  ...
</document>
```

Rysunek 2.6: Szablon pliku wynikowego branego pod uwagę w konkursie ICDAR - 2013

```
<?xml version="1.0" encoding="UTF-8"?>
<document filename='filename.jpg'>
  <table id='Table_1540517170416_3'>
    <Coords points="180,160 4354,160 4354,3287 180,3287"/>
    <cell id='TableCell_1540517477147_58' start-row='0' start-col='0' end-row='1' end-col='2'>
      <Coords points="180,160 177,456 614,456 615,163"/>
    </cell>
    ...
  </table>
  ...
</document>
```

Rysunek 2.7: Szablon pliku wynikowego branego pod uwagę w konkursie ICDAR - 2019

Cycle Name	KI (1/km)	Distance (mi)	Percent Fuel Savings			
			Improved Speed	Decreased Accel	Eliminate Stops	Decreased Idle
2012_2	3.30	1.3	5.9%	9.5%	29.2%	17.4%
2145_1	0.68	11.2	2.4%	0.1%	9.5%	2.7%
4234_1	0.59	58.7	8.5%	1.3%	8.5%	3.3%
2032_2	0.17	57.8	21.7%	0.3%	2.7%	1.2%
4171_1	0.07	173.9	58.1%	1.6%	2.1%	0.5%

Rysunek 2.8: Tabela o złożonej strukturze

Cycle Name	KI (1/km)	Distance (mi)	Percent Fuel Savings			
			Improved Speed	Decreased Accel	Eliminate Stops	Decreased Idle
2012_2	3.30	1.3	5.9%	9.5%	29.2%	17.4%
2145_1	0.68	11.2	2.4%	0.1%	9.5%	2.7%
4234_1	0.59	58.7	8.5%	1.3%	8.5%	3.3%
2032_2	0.17	57.8	21.7%	0.3%	2.7%	1.2%
4171_1	0.07	173.9	58.1%	1.6%	2.1%	0.5%

Rysunek 2.9: Konwersja tabeli (rys. 2.8) o złożonej strukturze do tabeli prostej wykorzystywana przez Camelot

2.5.1 Ewaluacja

Problem rozpoznawania struktury tabel jest skomplikowany, ale równie trudny jest problem ewaluacji wyników osiąganych przez rozwiązującego systemy. W uproszczeniu, celem ewaluacji jest to, by przy pomocy miar liczbowych ocenić stopień zgodności wyniku systemu z wynikiem oczekiwany i zrobić to w jak najszybszy i najbardziej zautomatyzowany sposób, zostawiając jednocześnie systemowi oceniającemu jak najmniejsze pole do interpretacji.

Z racji, że wyniki systemów często muszą w pewien sposób różnić się w zależności od zastosowania, utworzenie jednej, ustandaryzowanej formy ewaluacji wydaje się być trudne, a nawet w pewnym stopniu niepraktyczne. W literaturze można napotkać przykłady systemów, dla których zaproponowano nowatorską formę ewaluacji dostosowaną do konkretnego problemu lub technicznych szczegółów działania [7]. Często, zaproponowane formy ewaluacji nie przyjmują się w środowisku, przez co testy z ich wykorzystaniem nie są powielane przez inne systemy i w rezultacie wymierne zestawienie wyników nie jest możliwe.

Sytuację tą odmieniło w pewnym zakresie ogłoszenie konkursów rozpoznawania struktury ICDAR - 2013 i ICDAR - 2019. Forma konkursowa wymusiła wykorzystanie jednoznacznych,

zautomatyzowanych metod ewaluacji nadsyłanych prac. Do tego celu wybrano ogólną koncepcję przedstawioną w pracy Goebela [14].

Zakłada ona model tabeli z identyfikacją zarówno kraty, jak i układu tabeli (uwzględniającego komórki zespolone). Generowana jest jednowymiarowa lista sąsiedztw pomiędzy komórkami znajdującymi się w tabeli (z pominięciem komórek pustych). Wszystkie sąsiedztwa (połączenia) są klasyfikowane jako prawidłowe lub nieprawidłowe. W przypadku, gdy obaj sąsiedzi zgadzają się i kierunek sąsiedztwa również odpowiada ground - truth, połączenie jest klasyfikowane jako prawidłowe, w przeciwnym razie jako nieprawidłowe. Na podstawie dokonanego podziału, wyliczane są precyzja, czułość i f1 - score. Taką formę mierzenia błędów cechuje niezależność od absolutnej pozycji błędu w tabeli, przez co jest uodporniona na obecność nadmiarowych i fałszywie zespolonych komórek.

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(a) Prawidłowa struktura odpowiadająca ground - truth

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(b) Nieprawidłowo rozpoznana struktura z dodatkową kolumną

■ Prawidłowa relacja sąsiedztwa □ Nieprawidłowa relacja sąsiedztwa

Rysunek 2.10: Zasady oceniania połączeń w tabelach, rysunek pochodzący z pracy Gobela [14]

Przykładowo wyliczone miary ewaluacji dla tabeli z rysunku 2.10:

$$\text{Czułość} = \frac{\text{Poprawne relacje sąsiedztwa}}{\text{Wszystkie poprawne relacje sąsiedztwa}} = \frac{24}{31} = 77.4\% \quad (2.1)$$

$$\text{Precyzja} = \frac{\text{Poprawne relacje sąsiedztwa}}{\text{Wszystkie wykryte relacje sąsiedztwa}} = \frac{24}{28} = 85.7\% \quad (2.2)$$

Ta ogólna koncepcja została zaimplementowana w automatycznej formie przez organizatorów konkursu. W plikach XML, które miały być wynikiem działania systemów (rysunki 2.6 i 2.7) zapisane były potrzebne dane dotyczące zidentyfikowanych komórek i ich relacji sąsiedztwa. System oceniający wydobywał z pliku XML listę sąsiedztw pomiędzy komórkami i porównywał ją z listą sąsiedztw znajdującej się w ground - truth. By było to możliwe, komórki wynikowe musiały zostać skojarzone z komórkami zawartymi w odpowiedzi. W wersji z 2013 roku, kojarzenie

odbywało się poprzez unormowanie i proste porównanie zawartości tekstowej w komórkach. W 2019 roku, metoda ta została zmodyfikowana i porównywała komórki poprzez miarę zgodności nachodzenia na siebie okupowanych przez nie obszarów IOU [28] (po przekroczeniu ustalonego odgórnie progu zgodności, komórki były klasyfikowane jako tożsame). Jeśli co najmniej jednej z komórek z wyniku nie rozpoznano w ground - truth lub relacja sąsiedztwa była nieprawidłowa, połączenie klasyfikowano jako nieprawidłowe, w przeciwnym razie jako prawidłowe.

Pomimo, że obie implementacje w pewnym stopniu się przyjęły i są coraz częściej spotykane przy ewaluacjach nowo powstałych metod [25], [25], [27], [9], nie można powiedzieć, że nie mają wad. Fakt niesprawiedliwego oceniania metod opartych na wizji komputerowej przez implementację z 2013 roku został zaadresowany przez organizatorów i zaowocował modyfikacjami w 2019 roku. Nowa wersja ewaluacji wyeliminowała główne niedociągnięcie poprzedniczki, tym samym jednak lekceważąc kluczową część wyniku, czyli dobrą ekstrakcję zawartości tekstowej. Co więcej, generalnie niskie wyniki osiągane przez zestawiane metody na zbiorze konkursowym (nie przekraczające f1 - score na poziomie 0.44) pozwalają zastanawiać się nad tym, czy przyjęta metoda ewaluacji nie jest zbyt restrykcyjna.

2.5.2 Systemy TSR wybrane z literatury

TableNet [25]

Rozwiązanie głębokiego uczenia oparte na w pełni konwolucyjnych sieciach neuronowych (FCNN). W pre - processingu wykorzystuje moduł OCR do ekstrakcji słów i zamienia je na prostokaty pokolorowane zgodnie z typem danych, jaki się w nich znajduje (liczbowe, tekstowe). Bazą wykorzystanej sieci jest wytrenowana na publicznym zbiorze imagenet sieć VGG19. Architektura sieci została zmodyfikowana przez autorów zgodnie z ich intuicjami, tak, by na wyjściu otrzymane zostały dwie mapy sklasyfikowanych pikseli odpowiadających maskom wykrytych tabel i kolumn. Segmentacja wierszowa przeprowadzana jest z wykorzystaniem prostej heurystyki uwzględniającej zapełnienie kolumn. Model został dotrenowany na zbiorach ICDAR - 2013 i Marmot i oceniony na wcześniej odseparowanym podzbiorze ICDAR - 2013 zawierającym 34 dokumenty. Osiągnął wynik na poziomie 0.91 f1 - score.

CascadeTabNet [27]

Rozwiązanie głębokiego uczenia reprezentujące odmienne podejście oparte na rozpoznawaniu struktury poprzez detekcję komórek. W tym celu wykorzystywany został model Cascade Mask R - CNN rozbudowany na architekturze HRNet. Do treningu rozpoznawania struktury użyto podzbioru zbioru treningowego z konkursu ICDAR - 2019. Dodatkowo zastosowano dylatację i rozmywanie w celu augmentacji zbioru. System podczas detekcji tabeli równolegle klasyfikował ją

na obramowaną lub nieobramowaną. Zależnie od wyników klasyfikacji wybierana była jedna z dwóch ścieżek przetwarzania. Tryb tabel obramowanych wykorzystywał tradycyjne techniki segmentacji strukturalnej poprzez detekcję obramowań. Tryb tabel nieobramowanych wykorzystywał model głęboki w celu identyfikacji komórek i przeprowadzał wnioskowanie progowe na podstawie pozycji i nachodzenia na siebie komórek. Ewaluacja na części testowej zbioru ICDAR - 2019 wskazała na f1 - score na poziomie 0.44, co wystarczyło do wygrania tej edycji konkursu w tej kategorii.

Clusti [35]

Metoda bottom - up wykorzystująca moduł ekstrakcji tekstu CRAFT. Technika uczenia nienadzorowanego DBSCAN jest wykorzystywana zarówno przy wnioskowaniu segmentacji wierszowo - kolumnowej, jak i przy wstępny filtrowaniu szumów (poprzez odseparowanie klastrów pikseli na te o mniejszym i większym zagęszczeniu czarnego koloru). Parametry algorytmu segmentującego odpowiadające minimalnemu odstępowi pomiędzy klastrami i minimalnej liczbie słów w klastrze są dobierane algorytmicznie poprzez manualne zebranie wybranych statystyk z obrazu. Autor nie wyróżnia podziału na zbiory treningowy i testowy, podaje wyniki dla pełnych zbiorów ICDAR - 2013 i ICDAR - 2019 (odpowiednio 98.5 i 94.5 f1 - score). Metoda ewaluacji nie odpowiada dokładnie metodom stosowanym w konkursach.

Zuyev [36]

Metoda top - down oparta w dużym stopniu na tradycyjnych technikach wizji komputerowej. Silnikiem algorytmu było wydobycie połączonych komponentów z obrazu i rzutowanie ich na osie pionową i poziomą w celu przeprowadzenia wnioskowania dotyczącego kraty tabeli na dwóch sygnałach jednowymiarowych. Jak przekonuje autor publikacji, takie podejście wydaje się być naturalnym dla analizy pionowo i poziomo wyrównanych bloków tekstu występujących w tabelach i sprawdza się bardzo dobrze dla większości przypadków praktycznych. W pre - processingu użyto wybranych operacji morfologicznych, analizy połączonych komponentów, tradycyjnej metody eliminacji obramowań. Przy wnioskowaniu wykorzystane zostało również 2 - means clustering by poradzić sobie w jakimś stopniu z rzutowaniem scalonych komórek. Autor nie podał wyników konkretnej ewaluacji, ale twierdził, że metoda w 1997r. była zimportowana przez komercyjne oprogramowanie *FineReader OCR* (obecnie *ABBYY FineReader*), co może świadczyć o jej sprawdzonej jakości.

Metoda	Zbiór danych	Metoda ewaluacji	F - measure	Podejście
DeepDeSRT [29]	ICDAR-2013 dataset	ICDAR-2013	91.44	FCNN
Nurimnen [9]	ICDAR-2013 dataset	ICDAR-2013	94.6	System regułowy (wykorzystujący metadane)
Clusti [35]	ICDAR-2013 dataset	row - column overlapping measure	98.5	DBSCAN Clustering
TableNet [25]	ICDAR-2013 dataset	ICDAR-2013	90.1	FCNN
Clusti [35]	ICDAR-2019 dataset	row - column overlapping measure	94.5	DBSCAN Clustering
CascadeTabNet [27]	ICDAR-2019 dataset	ICDAR-2019	43.8	Cascade Mask R - CNN

Tablica 2.3: Zestawienie wybranych wyników z literatury pod kątem rozpoznawania struktury tabeli

Wszystkie przedstawione rozwiązania różnią się od siebie już w fundamentalnych założeniach, co pokazuje jak szeroka gama technik może być zastosowana w celu rozpoznania struktury tabeli. Reprezentanci popularnego obecnie podejścia opartego na CNN (TableNet, CascadeTabNet) potrafili osiągnąć stabilne rezultaty dla różnego typu danych wejściowych, a skalowanie algorytmów wydaje się być stosunkowo proste - wystarczy przygotować odpowiedni zestaw tabel treningowych dopasowanych do potrzeb aplikacji bez konieczności wprowadzania dużych zmian do modelów. Jest to też w pewnym stopniu ich ograniczeniem, ponieważ ze względu na duże zróżnicowanie w formatowaniu tabel, mogą nie być w stanie dobrze uogólniać sposobu wnioskowania, tym samym nie zwracając dobrych rezultatów dla stylów, z którymi się wcześniej nie spotkały.

Ciekawe wyniki przedstawił autor Clusti, którego rozwiązanie bottom - up osiągnęło wysokie wyniki segmentacji wierszowo - kolumnowej na stosunkowo trudnym zbiorze ICDAR2019. W opisie metody występują jednak pewne przesłanki, świadczące o tym, że mogłyby nie poradzić sobie z klasą tabel, w której wieloliniowe wiersze są położone w bardzo bliskim dystansie od siebie (taka klasa została zawarta w ramach zbioru danych, zaproponowanego przez niniejszą pracę magisterską).

Rozwiązanie Zuyeva można uznać za najbardziej intuicyjne, a fakt, że znalazło swoje praktyczne zastosowanie sprawia, że budzi szczególne zainteresowanie z punktu widzenia poruszanej problematyki. Szczegóły implementacji dotyczą jednak wielu niuansów, które razem z pozostałymi podejściami są bardziej wnikliwie poruszane w dalszej części pracy.

3. Przeprowadzone eksperymenty i badania

3.1 Usczegółowienie wymagań i przyjęte założenia

Jak wspomniano we wprowadzeniu, wybór niniejszego tematu był umotywowany potrzebą wdrożenia w prywatnym przedsiębiorstwie. Opracowany algorytm miał być częścią rozbudowanego systemu mającego na celu automatyzację procesu przyjmowania faktur zakupowych w systemie księgowym z równoległą aktualizacją stanu magazynowego.

Zadanie to, w wersji manualnej, polega na zlokalizowaniu w bazie magazynowej każdej pozycji znajdującej się w tabeli na podstawie informacji pobranej z jednej lub kilku kolumn (założnie od pozycji i struktury dokumentu wystawanego przez konkretnego dostawcę), wydobyciu informacji o ilości zakupionego towaru, jego cenie netto i brutto zakodowanych w osobnych kolumnach i zapisaniu tych wartości w bazie. Należy podkreślić, że pozycji na pojedynczej fakturze zakupowej może być wiele, podobnie jak faktur wprowadzanych do systemu każdego dnia, przez co zajęcie to jest bardzo żmudne i pracochłonne, a jego wtórna natura sugeruje możliwość automatyzacji i redukcji kosztów.

Warunkiem usprawnienia procesu jest korzystanie z faktur w łatwej do sparsowania, ustrukturyzowanej formie cyfrowej, takiej jak pliki CSV czy XML. Niektórzy dostawcy dostarczają takie pliki, choć często ich struktura znaczco różni się w zależności od dostawcy i każdy wymaga odpowiedniego przygotowania. Niestety, większość kontrahentów dostarcza jedynie cyfrowo wygenerowany plik PDF, z którego nie da się bezpośrednio wywnioskować struktury danych tabelarycznych. Zdarzają się również takie przypadki, w których wersja cyfrowa w ogóle nie jest dostarczana, a jedynie fizyczny, wydrukowany egzemplarz.

Wymienione trudności motywują opracowanie metody konwersji plików PDF i zdjęć na CSV wykorzystującej najbardziej uniwersalne przetwarzanie obrazów i technologię OCR. Metoda ta powinna być możliwie jak najbardziej dokładna na zaproponowanym zbiorze testowym pochodząącym z księgowości wspomnianego przedsiębiorstwa.

W początkowych założeniach, prace miały obejmować implementację i porównanie kilku pełnych algorytmów realizujących zarówno detekcję, jak i rozpoznawanie struktury tabel. Zrealizowanie tych założeń przez jedną osobę nie było możliwe, gdyż każdy z tych modułów jest

tematem obszernym i nadającym się na osobną pracę.

Moduł rozpoznawania struktury generuje najwięcej problemów przy obsłudze manualnej lub korygowaniu błędów systemów automatycznych przez co skupiono się właśnie na nim. Przy wdrożeniu miało zostać wykorzystane jedno z publicznie dostępnych rozwiązań detekcji [26], [2] wspierane ewentualną manualną korekcją. Ręczne zaznaczenie obszaru tabeli na zdjęciu nie jest zadaniem bardzo czasochłonnym i pozwala na pominięcie obszarów tabeli, które z punktu widzenia aplikacji docelowej nie są istotne.

W kolejnej iteracji, planowano zaimplementować i zestawić po jednym przedstawicielu metod top-down, bottom-up i głębokiego uczenia, jednak założenie to również okazało się zbyt ambitne. Nietrudno zauważyc, że prawie wszystkie rozwiązania z literatury były rozwijane przez zespoły kilkuosobowe. Implementacja dobrze działających algorytmów rozpoznawania tabeli opartych na analizie obrazu i ich wymierna ewaluacja jest zadaniem skomplikowanym i pracochłonnym. Wymaga dopracowanych działań pre- i post-processingowych, przemyślanych i często złożonych systemów wnioskujących, starannego przygotowania danych testowo - treningowych (szczególnie w podejściach opartych na uczeniu maszynowym), dodania i wykorzystania odpowiednich narzędzi ewaluacyjnych dopasowanych do wyników działania systemu.

By możliwe było praktyczne wykorzystanie efektów prac, dokładność opracowanego systemu musiała być wysoka, a ewentualne poprawki manualne plików wyjściowych nieduże i sporadyczne. Zdecydowano się opracować tylko jedną metodę, doszlifować ją do możliwie jak najlepszych wyników i zestawić z publicznie dostępnymi rozwiązaniami (bibliotekami, rozwiązaniami komercyjnymi) [2],[24] zaadaptowanymi do zaproponowanego zbioru testowego.

Określony musiał zostać model tabeli rozwijanego algorytmu, czyli założenia, w ramach których miał być realizowany. Podjęto decyzję o przyjęciu podejścia wykorzystującego jedynie analizę obrazu i metody OCR. W ramach niniejszej pracy, zbiór ewaluacyjny miał się składać tylko z dokumentów wygenerowanych cyfrowo, jednak w warunkach praktycznych, często wymagane jest również radzenie sobie ze skanami i zdjęciami dokumentów. Aby możliwe było dalsze rozwijanie opracowanej metody na analizę niedoskonałej jakości plików graficznych postanowiono nie korzystać z zakodowanych w plikach PDF metadanych. Nie zabroniono tego jednak rozwiązaniom zestawianym [2], choć z zaznaczeniem, że owe metadane dają pewną przewagę, szczególnie jeśli chodzi o niezawodną dokładność modułu ekstrakcji słów.

Zdecydowana większość tabel w rozpatrywanym zbiorze miała naturalnie prosty lub częściowo złożony układ. Z tego powodu, opracowana metoda miała ograniczyć się do rozpoznawania struktury poprzez identyfikację kraty (wyróżnienia kolumn i wierszy), z nadzieją, że wpłynie to pozytywnie na otrzymane rezultaty dla wspomnianych łatwiejszych przypadków.

Początkowo planowano w pełni zautomatyzować proces rozpoznawania struktury, jednak w miarę postępów prac zorientowano się że podejście półautomatyczne z przekazaniem kilku para-

metrów (na wzór Camelot [2]) może skutkować uproszczeniem problemu i co za tym idzie poprawą rezultatów.

Przyjęcie takiego założenia nie musiało odbijać się negatywnie na wartości wdrożeniowej - po jednorazowym dostrojeniu parametrów do danego stylu dokumentu (dostawcy), ten sam zestaw parametrów można wykorzystywać dla każdego innego dokumentu pochodzącego z tego samego źródła. Co więcej, w końcowych rezultatach, zdecydowana większość dokumentów była dobrze rozpoznawana z domyślnie ustawionym zbiorem wartości. Po obudowaniu takiego rozwiązania odpowiednim interfejsem, owe założenie, zdaniem autora, może być niezauważalne dla użytkownika końcowego, przynajmniej w kontekście konkretnego problemu, którego dotyczy praca.

3.2 Przygotowana baza dokumentów tabelarycznych

Przygotowane zostały pliki PDF zawierające 267 stron faktur zakupowych od 12 różnych kontrahentów. Każda strona przekonwertowana została na obraz .jpg w rozdzielcości 300 DPI. Przy pomocy darmowego narzędzia *makesense.ai* przygotowano adnotacje w formatach VOC XML, YOLO i CSV zawierające informacje o położeniu i typach tabel (w pełni obramowana, pionowo obramowana, poziomo obramowana, bez obramowań) znajdujących się na każdym z obrazów. Tabele wychodzące na więcej niż jedną stronę zostały rozdzielone by nie wprowadzać dodatkowych komplikacji.

Próbki w bazie charakteryzują się różnym stopniem złożoności. Zdecydowanie większa część z nich to tabele proste lub częściowo złożone (tak autor nazywa tabele, których złożoność jest widoczna jedynie w nagłówku lub stopce, a ich ciało ma strukturę idealnie prostą). Tabele, które posiadają zespolone komórki na całym swoim obszarze pochodzą z faktur tylko jednego dostawcy i są zdecydowanie najmniej liczne. Ogólne podsumowanie dotyczące bazy znajduje się w tabeli 3.1.

Ze względu na deficyt zadowalających automatycznych metod ewaluacji, zdecydowano się na wyodrębnienie z bazy podzbioru treningowo - testowego obejmującego wszystkie najbardziej zróżnicowane stylem tabele.

Proces selekcji tabel do odpowiednich podzbiorów przebiegał następująco:

1. Losowo wybrano po 3 dokumenty od każdego dostawcy. Wszystkie tabele znajdujące się w dokumentach trafiły do zbioru testowego i nie były brane pod uwagę aż do końcowej ewaluacji algorytmów.
2. Losowo dobrano po 2 kolejne dokumenty od każdego dostawcy. Wszystkie tabele znajdujące się w dokumentach trafiły do zbioru treningowego i były cyklicznie wykorzystywane do ewaluacji rozwijanego algorytmu.

3. Z pozostałych faktur dobrano kilka (głównie zawierających tabele częściowo obramowane), w których tabele, zdaniem autora, mogły sprawić największe problemy i dołączono je do zbioru treningowego.

W konsekwencji otrzymano zbalansowane zbiory, które mogły być w rozsądny czasie ewaluowane zarówno wizualnie, jak i w sposób bardziej systematyczny z wykorzystaniem przygotowanych narzędzi. Co więcej, tak wyselekcjonowany zbiór testowy dawał pewną dozę pewności, że zdobycie na nim dobrych wyników może zostać przełożone na skuteczne wdrożenie.

Opracowana metoda rozpoznawania tabeli była metodą regułową, wykorzystującą szereg transformacji, heurystyk, warunków i ustalonych eksperymentalnie parametrów (nie oparta na danych). Zbiór treningowy miał służyć jedynie do szybkiej ewaluacji modyfikowanego algorytmu bez wyliczania precyzyjnych statystyk. Z tego powodu, nie musiał zachowywać tradycyjnej przewagi objętościowej nad zbiorem testowym. Ręczne uzupełnienie zbioru treningowego najtrudniejszymi przypadkami ze zbioru głównego miało zagwarantować dobre działanie metody dla ogółu tabel i tym samym dobre wyniki w fazie testów.

Zidentyfikowano główne problemy towarzyszące analizie tabel ze zbioru. Były to:

- brak obramowań w jednej z płaszczyzn (rysunki 3.2, 3.3, 3.4, 3.5, 3.7)
- bardzo małe przestrzenie pomiędzy tekstem sąsiadujących komórek, które trudno było odróżnić od naturalnej przestrzeni pomiędzy słowami / liniami lub było to niemożliwe (rysunki 3.2, 3.7)
- przerywane obramowania (rysunek 3.7)
- złożone i częściowo złożone struktury tabel (rysunki 3.2, 3.3, 3.4, 3.6)
- zróżnicowanie czcionek i interlinii w dokumentach
- obecność mini - tabelek podsumowujących w strukturze tabeli (rysunki 3.2, 3.6)
- niejednolitość wizualnych wskazówek dotyczących segmentacji w przekroju pojedynczej tabeli (obramowanie jedynie nagłówka lub stopki, wyróżnienie niektórych regionów jedynie przez zmianę koloru tła, a innych przez obramowania) (rysunki 3.3, 3.5, 3.7)

Tabele 3.2, 3.3, 3.4 i 3.5 zawierają szczegółowe informacje o wydzielonych zbiorach treningowych i testowych z wyszczególnieniem złożoności tabel oraz informacji o wymiarach ich siatek.

Tabele w pełni obramowane	Tabele pionowo obramowane	Tabele poziomo obramowane	Tabele bez obramowania	Razem
233	40	90	0	363

Tablica 3.1: Liczebności tabel w zaproponowanej bazie ze względu na typ obramowania i złożoność

Złożoność \ Typ obramowania	Pełne	Pionowe	Poziome	Razem
Tabela prosta	20	1	10	31
Tabela częściowo złożona	5	9	0	14
Tabela złożona	0	0	4	4
Razem	25	10	14	49

Tablica 3.2: Liczebności tabel w zbiorze treningowym ze względu na typ obramowania i złożoność

Złożoność \ Typ obramowania	Pełne	Pionowe	Poziome	Razem
Tabela prosta	27	2	13	42
Tabela częściowo złożona	6	10	0	16
Tabela złożona	0	0	5	5
Razem	33	12	18	63

Tablica 3.3: Liczebności tabel w zbiorze testowym ze względu na typ obramowania i złożoność

	średnia liczba	odchylenie standardowe liczby	minimalna liczba	maksymalna liczba
kolumny	10.70	1.68	8	13
wiersze	17.38	10.49	2	38

Tablica 3.4: Statystyki dotyczące liczby kolumn i wierszy tabel w zbiorze treningowym

	średnia liczba	odchylenie standardowe liczby	minimalna liczba	maksymalna liczba
kolumny	10.61	1.49	8	13
wiersze	17.92	10.67	2	35

Tablica 3.5: Statystyki dotyczące liczby kolumn i wierszy tabel w zbiorze testowym

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

Lp	Nazwa towaru lub usługi	PKWIU	Ilość	Jedn. m.	Cena netto bez rabatu (zł)	Narzut Rabat (%)	Cena jedn. netto (zł)	Wartość netto (zł)	St %	Kwota podatku (zł)	Wartość brutto (zł)
1	Dlugopis S-FINE niebieski GWIAZDKI TO-059 TOMA dlk4490118		60	szt.	0,99	-19,00	0,80	48,00	23	11,04	59,04
2	Dlugopis S-FINE automatyczny niebieski TO-069 TOMA dlk4540118		30	szt.	1,15	-19,00	0,93	27,90	23	6,42	34,32
3	Nabojce PARKER (5sztuk) niebieskie 195038 4 nak0090055		6	szt.	7,64	-19,00	6,19	37,14	23	8,54	45,68
4	Dlugopis FLEXI niebieski PENNATE TT7038 dlk4960094		60	szt.	0,87	-19,00	0,70	42,00	23	9,66	51,66
5	Segregator A4/35mm 2 ringi czarny 14454 ESSELTE VIVIDA sek0470087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
6	Segregator A4/35mm 2 ringi żółty 14450 ESSELTE VIVIDA sek0430087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
7	Segregator A4/35mm 2 ringi zielony 14453 ESSELTE VIVIDA sek0460087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
8	Segregator A4/35mm 2 ringi bordowy 14456 ESSELTE VIVIDA sek0490087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
9	Segregator A4/35mm 2 ringi szary 14455 ESSELTE VIVIDA sek0480087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
10	Segregator A4/35mm 2 ringi czerwony 14451 ESSELTE VIVIDA sek0440087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
11	Segregator A4/35mm 2 ringi biały		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40

Rysunek 3.1: Tabela F-0034-10-G-21_0 pochodząca z autorskiego zbioru danych

Lp	Symbol	EAN	Symbol obcy	Opis	JM	Ilość	Cena PLN	Netto PLN	Vat		Brutto PLN
									%	PLN	
20	S 175 COC12	4007817048481		Kredki sześciokątne, kolekcja Comic, 12 kol, szt Staedtler		1	4,64	4,64	23%	1,07	5,71
21	S 965 14L NBK	4007817965023		Nożyczki Noris Club dla dzieci leworęcznych, 14 cm, blister, Staedtler		2	3,63	7,26	23%	1,67	8,93
22	S 550 55 BK	4007817185667		Cyrkiel szkolny, z uniwersalnym adapterem do piasków i ołówków, blister, Staedtler		2	9,55	19,10	23%	4,39	23,49
23	MG HACT1257 BK	6941255141893		Korektor w taśmie So Many Cats, 2 szt, 5m x 5mm, MG		10	4,53	45,30	23%	10,42	55,72
24	MG AS33V120	6941255143675		Karteczki samoprzyklepane So Many Cats; 7,6x7,6cm, 80ark, MG		20	2,38	47,60	23%	10,95	58,55
Razem								571,94	131,57	703,51	
Netto									Vat	Brutto	
W tym								571,94	23%	131,57	703,51

Rysunek 3.2: Tabela 2021_FA_MG01_2285_1 pochodząca z autorskiego zbioru danych

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

lp.	mag	kod kreskowy	nazwa towaru/usługi	kod CN /PKWIU	j.m	ilość	cena netto	wartość netto	% VAT	kwota vat	wartość brutto	masa (kg.)
1	A01	5902277274458	BBI LINIUKA FLEX 15CM B&B KIDS PASTEL	SZT	20	0.83	16.60	23%	3.82	20.42	0.300	
2	A01	5902277060006	BIOK RYSUNKOWY A4 20 80G	SZT	40	0.84	33.60	23%	7.73	41.33	5.440	
3	A01	5902277070012	BIOK TECHNICZNY A3 10 170G	SZT	120	1.95	234.00	23%	53.82	287.82	31.200	
4	A01	5902277070005	BIOK TECHNICZNY A4 10 170G	SZT	120	0.97	116.40	23%	26.77	143.17	15.600	
5	A01	5902277170828	BRULION A4 96# M 70G	SZT	10	4.46	44.60	23%	10.26	54.86	6.100	
6	A01	5902277171726	BRULION A4 96# M 70G	SZT	5	4.46	22.30	23%	5.13	27.43	3.050	
7	A01	5902277170804	BRULION A5 96# M 70G	SZT	10	2.35	23.50	23%	5.41	28.91	2.920	
8	A01	59022772394241	BRULION A5 96# M 70G METALLIC SATIN GOLD	SZT	5	2.96	14.80	23%	3.40	18.20	1.460	
9	A01	5902277170606	BRULION A6 96# M 70G	SZT	10	2.24	22.40	23%	5.13	27.55	1.450	
10	A01	5902277179760	DZIENNIK KORESPONDENCYJNY A4 96 70G	SZT	5	7.36	36.80	23%	8.46	45.26	2.700	
11	A01	5902277239518	IBB OLÓWEK B&B KIDS PASTEL	SZT	72	0.93	66.96	23%	15.40	82.36	0.432	
12	A01	59022772395262	IBB OLÓWEK ZE ZWIERZAKIEM B&B	SZT	36	1.90	68.40	23%	15.73	84.13	0.504	
13	A01	59022772746551	INT ZAKL. INDEKSOURCE FUNKY 200 12X45MM	SZT	3	2.72	8.16	23%	1.88	10.04	0.033	
14	A01	5902277274623	KOLOROWANKA 3 NAKI A4 16 DUŻO DO KOLOR.	58.11.13.0	SZT	10	2.06	20.60	5%	1.03	21.63	1.130
15	A01	5902277207418	KOLOR.Z.PP A5 100# SPJR.PO KR.BOKU Z G.B&B	SZT	5	5.81	29.05	23%	6.68	35.73	1.145	
16	A01	5902277171337	KOŁOZESZYT A5 160# M 70G Z PERF.	SZT	5	4.23	21.15	23%	4.86	26.01	1.750	
17	A01	5902277174437	KOSTKA PAP. 8.5x8.5x3.5CM BIAŁA KIEJONA	SZT	30	1.25	37.50	23%	8.63	46.13	8.190	
18	A01	5902277170569	KOSTKA PAP. 8.5x8.5x3.5CM KOLOR KIEJONA	SZT	60	1.63	97.80	23%	22.49	120.29	12.180	
19	A01	5902277178084	N-KOSTKA PAP. 8.5x8.5x3.5CM KOLOR NIEBIEK	SZT	10	1.63	16.30	23%	3.75	20.05	2.200	
20	A01	5902277244475	NSR KLEJ BROKATOWY MOSTER 6ml a'5	SZT	3	1.49	4.47	23%	1.03	5.50	0.195	
21	A01	5902277100009	PAPIER KOLOROWY A5 10 115G	SZT	10	0.67	6.70	23%	1.54	8.24	0.500	
22	A01	5902277275035	TDS DEJOPIS TODAYS JET LINE BLACK a'10	SZT	10	0.99	9.90	23%	2.28	12.18	0.100	
23	A01	5902277275042	TDS DEJOPIS TODAYS JET LINE BLUE a'10	SZT	10	0.99	9.90	23%	2.28	12.18	0.100	
24	A01	5902277170118	WKEAD DO SEG A4 50# KOL. M 70G	SZT	5	2.66	13.30	23%	3.06	16.36	1.175	
25	A01	5902277244453	YIN BROKAT SYPKI T7 A'6 PEARL	SZT	3	6.39	19.17	23%	4.41	23.58	0.315	
26	A01	5902277278203	YNT DEJOPIS ZELOWE 6SZT. NEON	SZT	2	6.40	12.80	23%	2.94	15.74	0.024	
27	A01	5902277278210	YNT DEJOPIS ZELOWE 6SZT. PASTEL	SZT	2	6.40	12.80	23%	2.94	15.74	0.024	
28	A01	5902277278197	YNT DEJOPIS ZELOWE 6SZT. METALIC	SZT	2	6.40	12.80	23%	2.94	15.74	0.220	
29	A01	5902277265531	ZESZEYT A4 60# M 70G PP	SZT	5	3.40	17.00	23%	3.91	20.91	1.500	
30	A01	5902277293862	N-ZESZEYT A4 60# M 70G HS KRAFT	SZT	5	3.15	15.75	23%	3.62	19.37	1.460	
31	A01	5902277175014	ZESZEYT A4 80# M 70G UV	SZT	5	3.14	15.70	23%	3.61	19.31	1.900	

K O N I E C S T R O N Y: 1

Rysunek 3.3: Tabela dok_int2_0 pochodząca z autorskiego zbioru danych

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

L.p.	Symbol Nazwa produktu	Ilość	J.m.	Rabat %	Cena netto	Kwota Netto	Podatek VAT %	Kwota Brutto
17	TC96/1 MB TABLICA KORKOWA 90X60CM W RAMIE DREWNIANEJ MEMOBOARDS [TC96/1 MB]	5903273024658	2 szt	25,00	15,26	30,52 P23	7,02	37,54
18	TC75/1 MB TABLICA KORKOWA 70X50CM W RAMIE DREWNIANEJ [TC75/1 MB]	5903273028113	1 szt	25,00	15,50	15,50 P23	3,57	19,07
19	TC85/1 MB TABLICA KORKOWA 80X50CM W RAMIE DREWNIANEJ [TC85/1 MB]	5903273032523	2 szt	25,00	16,34	32,68 P23	7,52	40,20
20	TM64ALC/1 MB TABLICA SUCHÓŚCIERALNO-MAGNETYCZNA 60X40CM W RAMIE ALUMINIOWEJ CLASSI [TM64ALC/1 MB]	5903273027161	1 szt	25,00	30,13	30,13 P23	6,93	37,06
21	TS96/1 MB TABLICA SUCHÓŚCIERALNA 90X60CM W RAMIE DREWNIANEJ [TS96/1 MB]	5903273037207	1 szt	25,00	21,12	21,12 P23	4,86	25,98
22	TS34/1 MB TABLICA SUCHÓŚCIERALNA 30X40CM W RAMIE DREWNIANEJ [TS34/1 MB]	5903273039676	1 szt	25,00	8,29	8,29 P23	1,91	10,20
23	154821 FC ZAKREŚLACZ 48 CZERWONY FABER-CASTELL [154821 FC]	4005401548218	10 szt	20,00	1,99	19,90 P23	4,58	24,48
24	154851 FC ZAKREŚLACZ 48 NIEBIESKI FABER-CASTELL [154851 FC]	4005401548515	10 szt	20,00	1,99	19,90 P23	4,58	24,48
25	154815 FC ZAKREŚLACZ 48 POMARAŃCZOWY FABER-CASTELL [154815 FC]	4005401548157	20 szt	20,00	1,99	39,80 P23	9,15	48,95
26	154828 FC ZAKREŚLACZ 48 RÓŻOWY FABER-CASTELL [154828 FC]	4005401548287	10 szt	20,00	1,99	19,90 P23	4,58	24,48
27	154863 FC ZAKREŚLACZ 48 ZIELONY FABER-CASTELL [154863 FC]	4005401548638	20 szt	20,00	1,99	39,80 P23	9,15	48,95
28	154807 FC ZAKREŚLACZ 48 ZŁOTY FABER-CASTELL [154807 FC]	4005401548072	20 szt	20,00	1,99	39,80 P23	9,15	48,95
29	7/5-BL ED TEXTMARKERY EDDING MINI 5 KOL. BLISTERER [7/5-BL ED]	4004764958795	1 kpl	20,00	6,39	6,39 P23	1,47	7,86
30	020-1760 NO ZSZYWACZ NOVUS E15 CZARNY Z PAKIETEM STARTOWYM ZSZYWEK No.10 DIN SUPER [020-1760 NO]	4009729048597	1 szt	20,00	7,48	7,48 P23	1,72	9,20
31	020-1474 NO ZSZYWACZ NOVUS C2 NIEBIESKI Z PAKIETEM STARTOWYM ZSZYWEK 24/6 DIN SUPER [020-1474 NO]	4009729020821	1 szt	20,00	18,39	18,39 P23	4,23	22,62
32	020-1472 NO ZSZYWACZ NOVUS C2 CZARNY Z PAKIETEM STARTOWYM ZSZYWEK 24/6 DIN SUPER [020-1472 NO]	4009729020876	4 szt	20,00	18,39	73,56 P23	16,92	90,48
33	020-1725 NO ZSZYWACZ NOVUS STABIL NIEBIESKI [020-1725 NO]	4009729046951	2 szt	20,00	12,47	24,94 P23	5,74	30,68
34	020-1287 NO ZSZYWACZ NOVUS STABIL CZARNY [020-1287 NO]	4009729009482	2 szt	20,00	12,47	24,94 P23	5,74	30,68

Rysunek 3.4: Tabela FSK_1803_0807_1 pochodzącej z autorskiego zbioru danych

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m ² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 220g/m ² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m ² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%
6	Folia laminacyjna PE arkusz A4 125 mic standard błysk antystatyczna Kod Indeksu: 320412	1	OP	29,50 PLN		29,50 PLN	29,50 PLN	23%
7	HEYKKA Zakreślacz klasyczny Linea jasnozielony, 10 szt./opk. Kod Indeksu: 611109	1	OP	21,00 PLN	35%	13,65 PLN	13,65 PLN	23%
8	HEYKKA Zakreślacz klasyczny Linea pastelowy żółty 10 szt./opk. Kod Indeksu: 611159	1	OP	21,00 PLN	35%	13,65 PLN	13,65 PLN	23%

Rysunek 3.5: Tabela FV210310097_0 pochodząca z autorskiego zbioru danych

Lp	INDEKS	NAZWA - OPIS TOWARU / USŁUGI	JM	ILOŚĆ	CENA NETTO	WARTOŚĆ POZYCJI BEZ PODATKU	PODATEK VAT %	KWOTA	WARTOŚĆ POZYCJI Z PODATKIEM
31	olk1190154	Ołówek automatyczny 0,5mm czarny TIKKY III ROTRING, 1904700	szt.	1.000	8.97	8.97	23	2.06	11.03
32	grk0410154 (11253)	Graffty do ołówków 0,5mm ROTRING S0312630	szt.	1.000	3.18	3.18	23	0.73	3.91
					Razem	972.42	x	223.66	1196.08
					W tym	972.42	23	223.66	1196.08

Rysunek 3.6: Tabela gd7.jrUT9w3LT_FV35021H2A2021_1 pochodząca z autorskiego zbioru danych

Lp.	Kod Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
25	291 Papier (jk) A4-100 czerwony (fluo)	5905824300853	2 opak.	4,9700	9,94	23%
26	504 Pap.fluo.pomarańcz samoprzyklejny A4-20	5905824800599	1 opak.	8,7700	8,77	23%
27	301 Pap.fluo.złoty samoprzyklejny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
28	476 Papier srebrny samoprzyklejny A4-20	5905824120628	2 opak.	14,5200	29,04	23%
29	648 Teczka (jk)A4/10 szk.pastel.z gumką lakier	5905824010431	5 opak.	10,4000	52,00	23%
30	129 Terminarz INFO A6 (kal.ks.tyg.)	5905824600984	10 szt.	2,9100	29,10	23%
31	139 Kalendarz VITO A4 chamois	5905824020669	4 szt.	29,3000	117,20	23%
32	550 Kalendarz menadżera A5	5905824900503	5 szt.	7,5900	37,95	23%
33	5066 Kalendarz B5	5905824501175	10 szt.	14,9000	149,00	23%
34	132 Kal.INFO A6 (mix)	5905824600175	5 szt.	4,9800	24,90	23%
35	131 Kal.INFO A5 (mix)	5905824400904	40 szt.	5,7900	231,60	23%
36	135 Terminarz SEVILLA A4 chamois	5905824801701	4 szt.	12,5000	50,00	23%

Rysunek 3.7: Tabela kr6_1 pochodząca z autorskiego zbioru danych

Pełna baza faktur nie mogła zostać opublikowana z uwagi na poufne informacje dotyczące stron transakcji. Autor może udostępnić zdjęcia wyciętych tabel należących do zbiorów na prośbę osób zainteresowanych.

3.3 Opracowana metoda rozpoznawania struktury tabel

3.3.1 Wybór podejścia

Przy wyborze podejścia i konkretnych algorytmów kierowano się przykładami z literatury i przeprowadzonymi wstępnie eksperymentami. W ostatnich latach najpopularniejszym podejściem do problemu rozpoznawania struktury tabeli było wykorzystanie konwolucyjnych sieci neuronowych. Wyniki działania znalezionych rozwiązań z literatury opartych na tych sieciach prezentowały się dobrze, dlatego postanowiono wypróbować je w pierwszej kolejności.

Odnaleziono dwie gotowe implementacje rozwiązań z literatury. CascadeTabNet zostało udostępnione przez autorów na stronie paperswithcode.com [6], odtwórcza implementacja rozwiązania TableNet została opublikowana w jednym z artykułów z portalu medium.com [26].

Sprawdzono działanie CascadeTabNet na wybranych tabelach ze zbioru treningowego. Uzykane wyniki segmentacji (detekcji komórek) pokazano na rysunkach 3.8, 3.9 i 3.10.

Lp	Nazwa towaru lub usługi	PKWIU	Ilość	Jedn. m.	Cena netto bez rabatu (zł)	Narzut (%)	Cena jedn. netto (zł)	Wartość netto (zł)	St %	Kwoty godzidły (zł)	Wartość brutto (zł)
1	Długopis S-FINE niebieski 14453 TOMA 14453 TONKA		60	szt.	0,99	-19,00	0,80	48,00	23	11,04	59,04
2	Długopis S-FINE automatyczny niebieski TO-069 TOMA 14453 TOMA 14453 TONKA		30	szt.	1,15	-19,00	0,93	27,90	23	6,42	34,32
3	Nabojce PARKER (5sztuk) niebieskie 195038 4 nak0090055		6	szt.	7,64	-19,00	6,19	37,14	23	8,54	45,68
4	Długopis FLEXI niebieski PENNATE TT7038 14454 960094		60	szt.	0,87	-19,00	0,70	42,00	23	9,66	51,66
5	Segregator A4/35mm 2 ringi czarny 14454 ESSELTE VIVIDA sek0470087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
6	Segregator A4/35mm 2 ringi szary 14450 ESSELTE VIVIDA sek0430087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
7	Segregator A4/35mm 2 ringi zielony 14453 ESSELTE VIVIDA sek0460087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
8	Segregator A4/35mm 2 ringi bordowy 14456 ESSELTE VIVIDA sek0490087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
9	Segregator A4/35mm 2 ringi szary 14455 ESSELTE VIVIDA sek0470087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
10	Segregator A4/35mm 2 ringi szary 14451 ESSELTE VIVIDA sek0470087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
11	Segregator A4/35mm 2 ringi szary 14452 ESSELTE VIVIDA sek0470087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40

Rysunek 3.8: Wyniki detekcji komórek osiągnięte przez niemodyfikowany model CascadeTabNet na tabeli F-0034-10-G-21_0 ze zbioru treningowego

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

Lp.	Symbol	Nazwa produktu	Ilość	J.m.	Netto % PLN	Cena netto	Kwota Netto % Kwota	Podatek VAT % Kwota	Kwota Brutto
17	TC96/1 MB	5903273024658 TABLICA KORKOWA 90X60CM W RAMIE DREWNIANEJ MEMOBOARDS [TC96/1 MB]	2 szt	25,00	15,26	30,52 P23	7,02	37,54	
18	TC75/1 MB	5903273028113 TABLICA KORKOWA 70X50CM W RAMIE DREWNIANEJ [TC75/1 MB]	1 szt	25,00	15,00	30,00 P23	3,57	33,07	
19	TC85/1 MB	5903273032523 TABLICA KORKOWA 80X30CM W RAMIE DREWNIANEJ [TC85/1 MB]	2 szt	25,00	16,04	32,08 P23	7,02	39,10	
20	TM64ALC/1 MB	5903273027161 TABLICA SUCHOŚCIERALNA-MAGNETYCZNA 60X40CM W RAMIE ALUMINIOWEJ CLASSI [TM64ALC/1 MB]	1 szt	25,00	30,13	30,13 P23	6,83	37,06	
21	TS96/1 MB	5903273037207 TABLICA SUCHOŚCIERALNA 90X60CM W RAMIE DREWNIANEJ [TS96/1 MB]	1 szt	25,00	21,12	21,12 P23	4,86	25,98	
22	TS94/1 MB	5903273039676 TABLICA SUCHOŚCIERALNA 30X40CM W RAMIE DREWNIANEJ [TS94/1 MB]	1 szt	25,00	8,29	8,29 P23	1,91	10,20	
23	154821 FC	4005401548218 ZAKRESŁACZ 48 CZERWONY FABER-CASTELL [154821 FC]	10 szt	20,00	1,99	19,90 P23	4,58	24,48	
24	154851 FC	4005401548515 ZAKRESŁACZ 48 NIEBIESKI FABER-CASTELL [154851 FC]	10 szt	20,00	1,99	19,90 P23	4,58	24,48	
25	154815 FC	4005401548157 ZAKRESŁACZ 48 POMARAŃCZOWY FABER-CASTELL [154815 FC]	20 szt	20,00	3,98	39,80 P23	9,15	49,95	
26	154828 FC	4005401548287 ZAKRESŁACZ 48 RÓŻOWY FABER-CASTELL [154828 FC]	10 szt	20,00	1,99	19,90 P23	4,58	24,48	
27	154863 FC	4005401548638 ZAKRESŁACZ 48 ZIELONY FABER-CASTELL [154863 FC]	20 szt	20,00	1,99	19,90 P23	9,15	48,95	
28	154802 FC	4005401548072 ZAKRESŁACZ 48 ZŁOTY FABER-CASTELL [154802 FC]	20 szt	20,00	1,99	19,90 P23	9,15	48,95	
29	TM-BL ED	4004764958795 TEXTMARKERY EDDING MINI 5 KOL. BLISTERER [7/5-BL ED]	1 kpl	20,00	0,99	0,99 P23	1,47	7,86	
30	020-1760 NO	4009729048597 ZSZYWACZ NOVUS E15 CZARNY Z PAKIETEM STARTOWYM ZSZYWEK No.10 DIN SUPER [020-1760 NO]	1 szt	20,00	7,48	7,48 P23	1,72	9,20	
31	020-1474 NO	4009729020821 ZSZYWACZ NOVUS C2 NIEBIESKI Z PAKIETEM STARTOWYM ZSZYWEK 24/6 DIN SUPER [020-1474 NO]	1 szt	20,00	18,39	18,39 P23	4,23	22,62	
32	020-1472 NO	4009729020876 ZSZYWACZ NOVUS C2 CZARNY Z PAKIETEM STARTOWYM ZSZYWEK 24/6 DIN SUPER [020-1472 NO]	4 szt	20,00	18,39	73,56 P23	16,92	90,48	
33	020-1725 NO	4009729046951 ZSZYWACZ NOVUS STABIL NIEBIESKI [020-1725 NO]	2 szt	20,00	12,47	24,94 P23	5,74	30,68	
34	020-1287 NO	4009729009482 ZSZYWACZ NOVUS STABIL CZARNY [020-1287 NO]	2 szt	20,00	12,47	24,94 P23	5,74	30,68	

Rysunek 3.9: Wyniki detekcji komórek osiągnięte przez niemodyfikowany model CascadeTabNet na tabeli FSK_1803_0807_1 ze zbioru treningowego

Lp	Symbol	EAN	Symbol czytnicy	Opis	JM	Ilość	Cena PLN	Netto PLN	Mat %	Brutto PLN
1	S 8323 BK2	4007817832059		Marker metaliczny, okragła końcówka, M, złoty i srebrny, blister, Staedtler	szt	1	8,10	8,10	23%	9,96
2	HA 7370 0075-0	5905130000074		Farba akrylowa 75 ml, biały tytanowa, Happy Color	szt	12	4,05	48,60	23%	59,40
3	HA 7370 0075-56	5905130000250		Farba akrylowa 75ml, ciemna oliwka, Happy Color	szt	2	4,05	8,06	23%	9,91
4	HA 7370 0075-75	5905130000260		Farba akrylowa 75ml, ciemnobrązowy, Happy Color	szt	4	4,05	16,12	23%	19,83
5	HA 7370 0075-9	5905130000357		Farba akrylowa 75ml, czarny, Happy Color	szt	6	4,05	24,18	23%	29,24
6	HA 7370 0075-201	5905130000979		Farba akrylowa 75ml, czerwony fluo, Happy Color	szt	1	4,05	4,05	23%	5,07
7	HA 7370 0075-8	5905130000264		Farba akrylowa 75ml, fioletowy, Happy Color	szt	2	4,05	8,06	23%	9,91
8	HA 7370 0075-435	5905130001250		Farba akrylowa 75 ml,kość słoniowa, Happy Color	szt	5	4,05	20,15	23%	24,70
9	HA 7370 0075-221	5905130000560		Farba akrylowa 75ml, różowy fluo, Happy Color	szt	1	4,05	4,45	23%	5,47
10	HA 7370 0075-80	5905130000815		Farba akrylowa 75ml, szary, Happy Color	szt	3	4,05	12,09	23%	14,87
11	HA 7370 0075-505	5905130000973		Farba akrylowa 75ml , zielony pastelowy, Happy Color	szt	1	4,05	4,03	23%	5,06
12	HA 7370 0075-101	5905130000448		Farba akrylowa 75ml, żółty fluo, Happy Color	szt	1	4,05	4,45	23%	5,47
13	HA AKR67K5-3	4007817812298		Wkład do długopisu usuwalnego, Standard A, 0.5mm, niebieski, 3 szt w etui, Happy Color	szt	10	5,05	50,50	23%	60,31
14	S 364 WP A	4007817841628		Zakreslacz Classic, 8 szt., w etui, 2 złoty, Staedtler	szt	1	21,03	21,03	23%	25,99
15	HA 7370 0075-7	5905130000033		Farba akrylowa 75ml, brązowy, Happy Color	szt	2	4,05	8,06	23%	9,91
16	HA 7370 0075-56	5905130000255		Farba akrylowa 75ml, ciemna oliwka, Happy Color	szt	2	4,05	8,06	23%	9,91
17	HA 7370 0075-75	5905130000262		Farba akrylowa 75ml, ciemnobrązowy, Happy Color	szt	4	4,05	16,12	23%	19,83
18	HA 7370 0075-25	5905130000222		Farba akrylowa 75ml, ciemnoróżowy, Happy Color	szt	1	4,05	4,03	23%	5,06
19	HA 7370 0075-9	5905130000350		Farba akrylowa 75ml, czarny, Happy Color	szt	4	4,05	16,12	23%	19,83

Rysunek 3.10: Wyniki detekcji komórek osiągnięte przez niemodyfikowany model CascadeTabNet na tabeli 2020_FA_MG01_14053_0 ze zbioru treningowego

Algorytm w założeniu miał wykrywać komórki tabeli, następnie na ich podstawie wnioskować

wać o strukturze tabeli. Testy pokazały, że już detekcja komórek wyraźnie nie spełniała oczekiwania - większość komórek była wykrywana błędnie lub nie była wykrywana wcale.

Oczywistym stało się, że zastosowanie modelu głębokich sieci neuronowych bez dotrenowania go na danych ze zbioru treningowego nie miało szansy powodzenia. Potwierdziło to popularne w literaturze spostrzeżenie, że generalnie tabele występujące w praktyce są mocno zróżnicowane i dopracowanie algorytmu na jednym ich zbiorze wcale nie musi gwarantować porównywalnych wyników na innych.

Przeprowadzenie treningu na własnych danych wymagało przygotowania odpowiedniego ground-truth dostosowanego do potrzeb modelu. To zadanie z kolei było skomplikowane, ponieważ wymagało:

- przygotowania narzędzi wspomagającego zaznaczanie odpowiednich regionów i zapisywania ich do formatu wymaganego przez model
- manualnego przygotowania adnotacji
- znalezienia sposobu na poradzenie sobie z dużym zagęszczeniem komórek i występowaniem niejednoznaczności w niektórych tabelach (np. tabela 3.4)

Dopiero po wykonaniu tych kroków, wprowadzanie modyfikacji do czynności pre- i post - przeształceniowych czy też modułu wnioskującego miałyby sens.

Ze względu na złożoność zadania i niepewność czy końcowy efekt będzie wystarczająco dobry, by go wdrożyć, postanowiono odłożyć pracę nad modelem sieci CNN i skupić się na rozwiązańach alternatywnych.

Znaleziona implementacja TableNet nie zawierała żadnego modułu segmentacji wierszowej. Nie posiadając w tamtym momencie doświadczenia w implementowaniu tego typu modułów od podstaw, zrezygnowano z rozwijania tej metody bez wykonywania żadnych wstępnych testów.

W kolejnym etapie prac rozpatrzone wykorzystanie metody bottom - up. Nie odnaleziono gotowych, przykładowych implementacji, ale zebrane artykuły ([22], [27]) zawierały dużo informacji, jak taką metodę można opracować od podstaw.

Warunkiem dobrego działania algorytmów bottom - up jest dopracowany moduł ekstrakcji słów z tabeli. W przypadku metody opartej na analizie obrazu jest to moduł optycznego rozpoznawania tekstu (OCR).

Postanowiono przetestować działanie konkretnego modułu OCR do ekstrakcji słów na wybranych tabelach ze zbioru treningowego. Modułem tym był *pytesseract* [23], wrapper silnika *Tesseract-OCR Engine* [30] w języku python. Bez stosowania żadnego przetwarzania wstępnego użyto ekstraktora na wyciętych obrazach tabel w trybie segmentacji strony służącym do rozpoznawania pionowo wyrównanych bloków tekstowych (dedykowanym do ekstrakcji tekstu z tabel). Wyniki zostały pokazane na rysunkach 3.11, 3.12, 3.13.

O ile wyniki osiągnięte na tabelach 3.11 i 3.13 można uznać za obiecujące i po dodatkowym pre - processingu (binaryzacja, szkieletyzacja, wyeliminowanie obramowań) mogłyby być niemal perfekcyjne, o tyle na tabeli z rysunku 3.12 można dostrzec wiele pomyłek i ciężko przewidzieć ile pracy wymagałoby doprowadzenie rezultatów do akceptowalnego stanu. Pamiętając, że stylów tabel w zbiorze jest więcej i że warunkiem przejścia do kolejnych etapów implementacji jest doprowadzenie modułu OCR do bardzo dobrych wyników dla każdego z nich, zdecydowano się tymczasowo zrezygnować z kontynuowania tego podejścia i rozpatrzyć pozostałe możliwości.

Rysunek 3.11: Wyniki detekcji słów osiągnięte przez moduł *pytesseract* na przykładowej tabeli F-0034-10-G-21_0 ze zbioru treningowego

	Indeks	Nazwa pełna	Ilość	NETTO	BRUTTO	Wartość NETTO	Stawka VAT
24	196774	LUPA ZELENKOWA 10X	1	3,51	3,51	2,85	23%
	5907690834654	KREDKI OL.6 KOLORINO NEON+TEMP.	1	14,92	14,92	14,92	23%
26	FPPKA1	FARTUSZEK DO PRAC PLASTYCZNYCH KIDEA	1	0,60	0,60	0,54	23%
	436281	BRELON ZELENKOZIEMIA 10cm	1	2,90	2,90	2,90	23%
	5907604608883	LUPA ZELENKOWA 10X	1 szt	2,90	2,90	2,90	23%
9	110850	WATERMAN JABOKE SLUGH FARNESIANA	1	8,42	8,42	8,42	23%
30	CR1018	KLEJ MAGIC TURA 20	1	2,15	2,15	2,15	23%
	HENRY-2475	PODKLADKA DLA BURKO	1	3,39	3,39	3,39	23%
32	HENRY-3693	PODKLADKA DLA BURKO	1 szt	3,39	3,39	3,39	23%
33	HENRY-3709	PODKLADKA DLA BURKO NAKI DROGOWE	1	3,39	3,39	3,39	23%
34	HENRY-1478	PODKLADKA DLA BURKO RY	1 szt	3,39	3,39	3,39	23%
35	PO-A3X-8191-XXX	Podkład laminowany LITTLE FRIEND	1	2,13	2,13	2,13	23%
36	ANT-NA65	NOTES ANTRA LOGO	1	9,59	9,59	9,59	23%
37	ANT-NA6	NOTES ANTRA A6 mix	1	5,62	5,62	5,62	23%
38	ANT-NA6	NOTES ANTRA A6 mix	1	5,62	5,62	5,62	23%
39	FTT010	DŁUG FLEXI ZIELONY ALU	1	0,81	0,81	13,20	23%
40	6435T	KLEJ OSTK. DOCY 50 ML	1 szt	29,42	29,42	29,42	23%
41	TT16445	trans. www.sklep.drogeria-polska.pl	1	16,26	16,26	16,26	23%
42	TO-069Z1	DŁUG GWIAZDKA 1-FINE SZWAJCARSKI WKD	1	0,17	0,17	0,17	23%

Rysunek 3.12: Wyniki detekcji słów osiągnięte przez moduł *pytesseract* na tabeli FA_12-10-18-Fwz-HS_Oryginal_1_16_1 ze zbioru treningowego

Rysunek 3.13: Wyniki detekcji słów osiągnięte przez moduł *pytesseract* na tabeli 2020_FA_MG01_14053_0 ze zbioru treningowego

Bardzo popularną, elementarną metodą segmentacji strukturalnej tabel jest metoda top-down wykorzystująca operacje morfologiczne do ekstrakcji linii i wnioskowania o układzie tabeli na ich podstawie [1]. Tabele w pełni obramowane występują w praktyce stosunkowo często i metoda ta

radzi sobie na tyle dobrze, że znalazła swoje zastosowanie w bardziej zaawansowanych systemach [20], [2].

Niestety, rzadko zdarza się, aby narzędzie musiało radzić sobie jedynie z tabelami w pełni obramowanymi. Również w zaproponowanym zbiorze testowo - treningowym, niemal 50% stanowią tabele obramowane tylko na jednej osi.

Dobrym pomysłem wydawało się opracowanie metody top - down w pełni wykorzystującej informacje o obramowaniach i rozbudowanie jej o wnioskowanie również na podstawie innych wizualnych przesłanek. W takim podejściu, wykorzystanie tylko operacji morfologicznych nie było wystarczające.

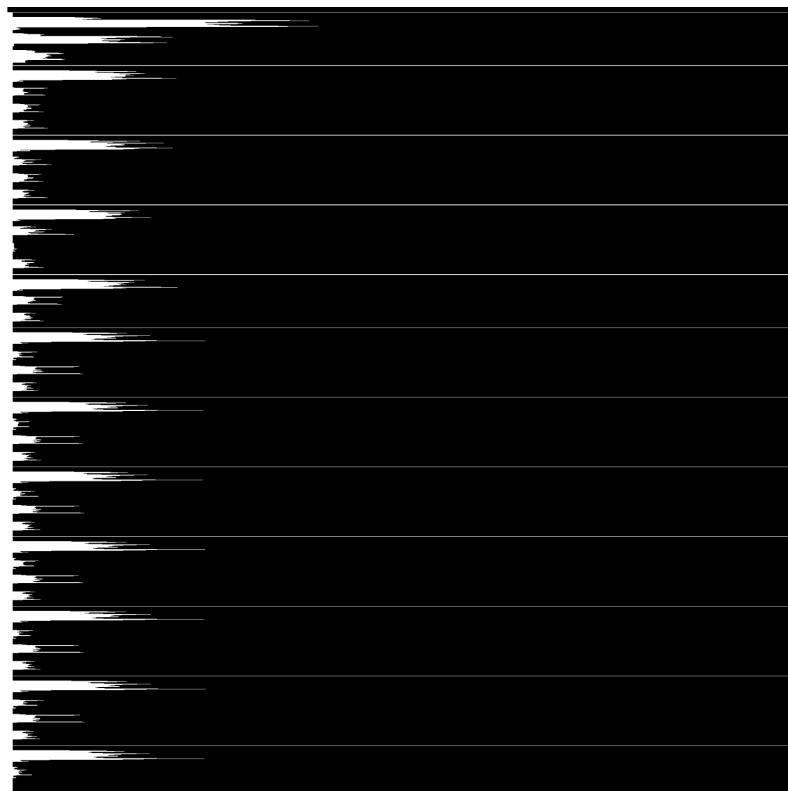
W systemach top - down, pewną popularność zyskało wykorzystanie projekcji profilowej [17], [36]. Szczególnie ciekawa i dobrze opisana okazała się metoda przedstawiona przez Zuyeva [36].

Przeprowadzono eksperyment, w celu zweryfikowania zasadności wykorzystania tego podejścia dla tabel należących do analizowanego zbioru przygotowanych faktur.

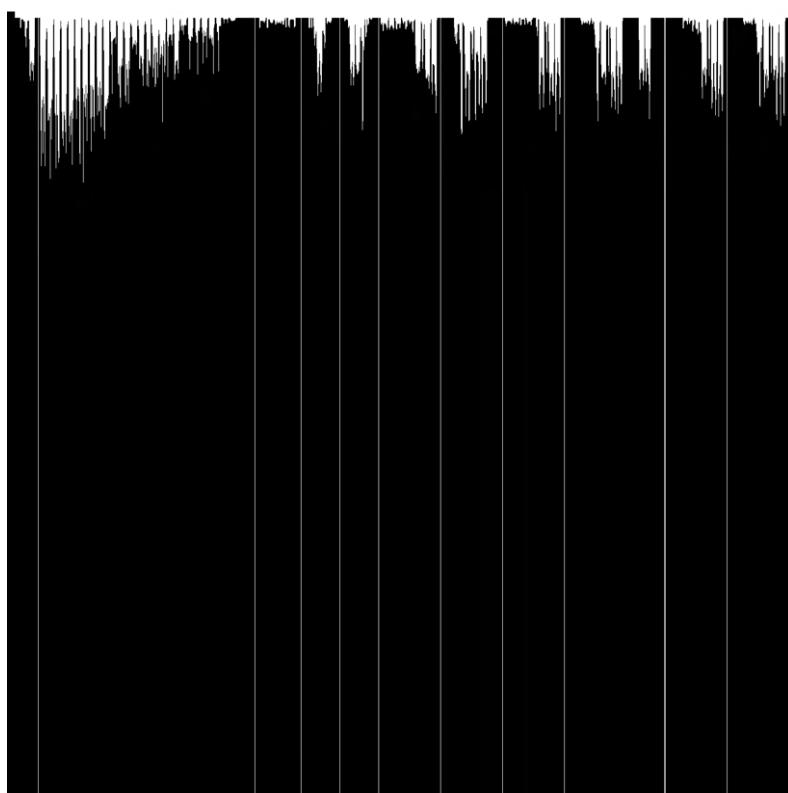
Sporządzono projekcje kilku tabel ze zbioru treningowego na osie pionową i poziomą. Rzurowanie pikseli zrealizowane zostało w sposób surowy, bez dodatkowego przetwarzania (w przeciwieństwie do metody Zuyeva). Wyniki zostały zaprezentowane na rysunkach 3.14 - 3.22.

Lp	Nazwa towaru lub usługi	PKWIU	Ilość	Jedn. m.	Cena netto bez rabatu (zł)	Narzut Rabat (%)	Cena jedn. netto (zł)	Wartość netto (zł)	St %	Kwota podatku (zł)	Wartość brutto (zł)
1	Diugopis S-FINE niebieski GNAZDZKI TO-059 TOMA dlk4490118		60	szt.	0,99	-19,00	0,80	48,00	23	11,04	59,04
2	Diugopis S-FINE automatyczny niebieski TO-069 TOMA dlk4540118		30	szt.	1,15	-19,00	0,93	27,90	23	6,42	34,32
3	Naboj Parker (5sztuk) niebieskie 195038 4 nak0090055		6	szt.	7,64	-19,00	6,19	37,14	23	8,54	45,68
4	Diugopis FLEXI niebieski PENNATE T77038 dlk4960094		60	szt.	0,87	-19,00	0,70	42,00	23	9,66	51,66
5	Segregator A4/35mm 2 ringi czarny 14454 ESSELTE VIVIDA sek0470087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
6	Segregator A4/35mm 2 ringi żółty 14450 ESSELTE VIVIDA sek0430087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
7	Segregator A4/35mm 2 ringi zielony 14453 ESSELTE VIVIDA sek0460087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
8	Segregator A4/35mm 2 ringi bordowy 14456 ESSELTE VIVIDA sek0490087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
9	Segregator A4/35mm 2 ringi szary 14455 ESSELTE VIVIDA sek0480087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
10	Segregator A4/35mm 2 ringi czerwony 14451 ESSELTE VIVIDA sek0440087		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40
11	Segregator A4/35mm 2 ringi biały		2	szt.	7,73	-19,00	6,26	12,52	23	2,88	15,40

Rysunek 3.14: Tabela F-0034-10-G-21_0 ze zbioru treningowego po binaryzacji



Rysunek 3.15: Projekcja pikseli tabeli z rys. 3.14 na oś poziomą



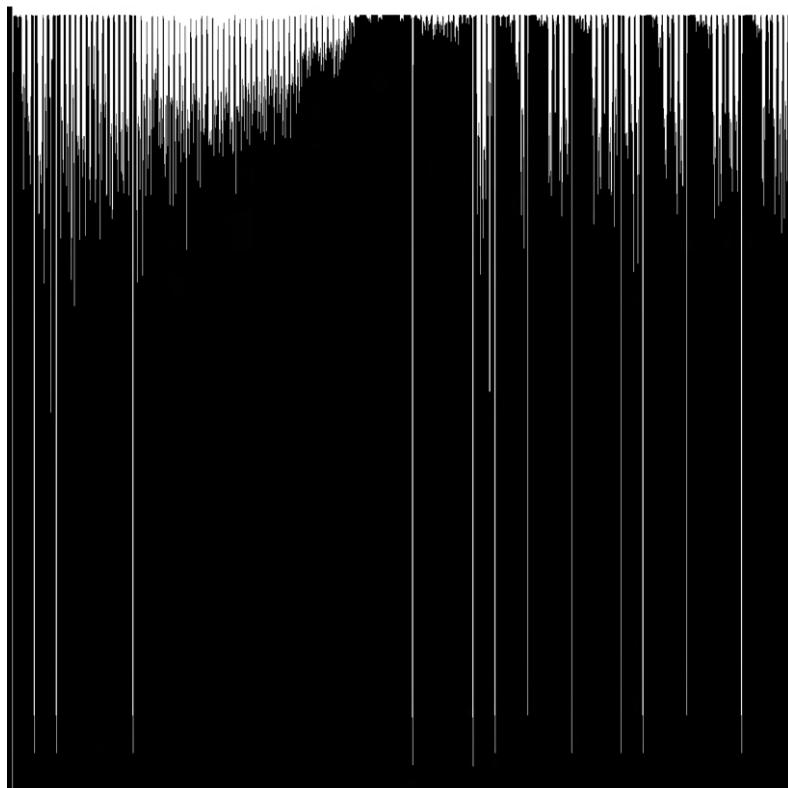
Rysunek 3.16: Projekcja pikseli tabeli z rys. 3.14 na oś pionową

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

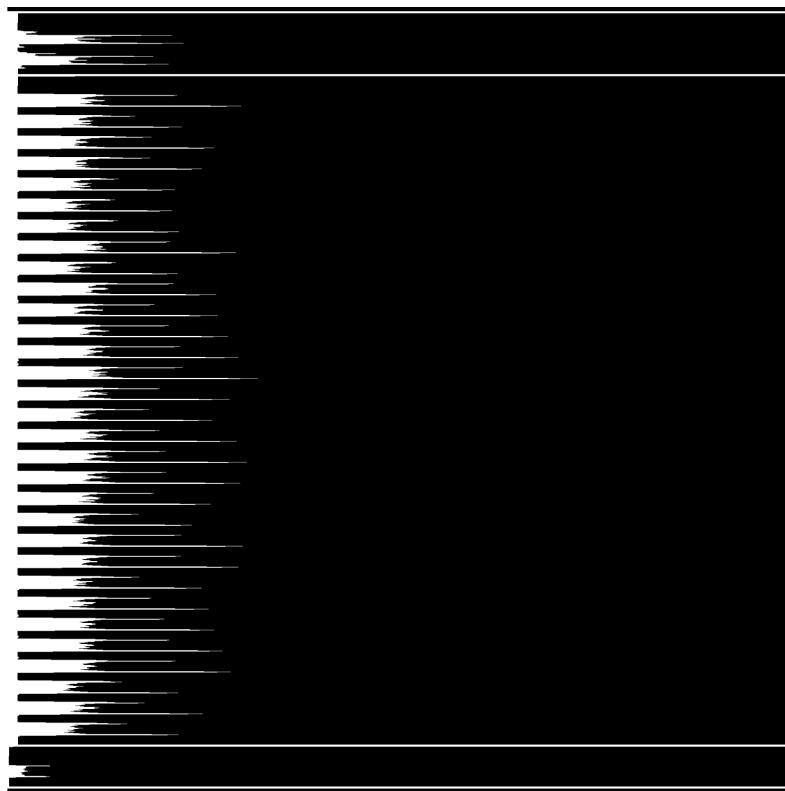
lp.	mag.	kod kreskowy	nazwa towaru/usługi	kod CN /PKNU	j.m	ilosc	cena netto	wartosc netto	% VAT	kwota vat	wartosc brutto	masa (kg.)
1	A01	590227724458	BBI LINIUKA FLEX 15CM B&B KIDS PASTEL	SZT	20	0.83	16.60	23%	3.82	20.42	0.300	
2	A01	5902277060006	BLOK RYSUNKOWY A4 20 80G	SZT	40	0.84	31.60	23%	7.73	41.33	5.440	
3	A01	5902277070012	BLOK TECHNICZNY A3 10 170G	SZT	120	1.95	234.00	23%	53.82	287.82	31.200	
4	A01	5902277070005	BLOK TECHNICZNY A4 10 170G	SZT	120	0.97	116.40	23%	26.77	143.17	15.600	
5	A01	5902277170828	BRULION A4 96# M 70G	SZT	10	4.46	44.60	23%	10.26	54.86	6.100	
6	A01	5902277171726	BRULION A4 96# M 70G	SZT	5	4.46	22.30	23%	5.13	27.43	3.050	
7	A01	5902277170824	BRULION A5 96# M 70G	SZT	10	2.35	23.50	23%	5.41	28.31	2.920	
8	A01	5902277294241	BRULION A5 96# M 70G METALLIC SATIN GOLD	SZT	5	2.96	14.80	23%	3.40	18.20	1.460	
9	A01	5902277170656	BRULION A6 96# M 70G	SZT	10	2.24	22.40	23%	5.15	27.55	1.450	
10	A01	5902277179760	DZIENNIK KORESPONDENCYJNY A4 96 70G	SZT	5	7.36	36.80	23%	8.46	45.26	2.700	
11	A01	5902277295118	IBB OLÓMEK B&B KIDS PASTEL	SZT	72	0.93	66.36	23%	15.40	82.36	0.432	
12	A01	5902277295242	IBB OLÓMEK ZE ZWIERZAKIEM B&B	SZT	36	1.90	68.40	23%	15.73	84.13	0.504	
13	A01	5902277276551	INT ZAKL.INDeksujace FUNKY 200 12X45MM	SZT	3	2.12	8.16	23%	1.88	10.04	0.033	
14	A01	5902277274625	KOLOROWANKA Z NAKL.A4 16 DÜŻO DO KOLOR.	SZT	10	2.06	20.60	5%	1.03	21.63	1.130	
15	A01	5902277207418	KOŁOZ.PP A5 100g SPŁ.PO KR.BOXU Z G.B&B	SZT	5	5.81	29.05	23%	6.68	35.73	1.145	
16	A01	5902277171337	KOŁOZESZEYI A5 160# M 70G Z PERF.	SZT	5	4.23	21.15	23%	4.86	26.01	1.750	
17	A01	5902277174437	KOSTKA PAP.8.5x8.5x3.5CM BIAŁA KLEJONA	SZT	30	1.25	37.50	23%	8.63	46.13	8.190	
18	A01	5902277170569	KOSTKA PAP.8.5x8.5x3.5CM KOLOR KLEJONA	SZT	60	1.63	97.80	23%	22.49	120.29	12.180	
19	A01	5902277178054	N-KOSTKA PAP.8.5x8.5x3.5CM KOLOR NIEKLEJ	SZT	10	1.63	16.30	23%	3.75	20.05	2.200	
20	A01	5902277244475	NSR KLEJ BROKATOWY NOSTER fml a'5	SZT	3	1.49	4.47	23%	1.03	5.50	0.195	
21	A01	5902277100009	PAPIER KOLOROWY A5 10 115G	SZT	10	0.67	6.70	23%	1.54	8.24	0.500	
22	A01	5902277275035	TDS DUUGOPIS TODAYS JET LINE BLACK a'10	SZT	10	0.99	9.90	23%	2.28	12.18	0.100	
23	A01	5902277275042	TDS DUUGOPIS TODAYS JET LINE BLUE a'10	SZT	10	0.99	9.90	23%	2.28	12.18	0.100	
24	A01	5902277170118	WŁAD DO SEC A4 50# KOL.M 70G	SZT	5	2.66	13.30	23%	3.06	16.36	1.175	
25	A01	5902277244543	YNA BROKAI SYEKI 7G A4 6# PEARL	SZT	3	6.39	19.17	23%	4.41	23.58	0.315	
26	A01	5902277278203	YNT DUUGOPIS ŻELOWE GSZT. NEON	SZT	2	6.40	12.80	23%	2.94	15.74	0.024	
27	A01	5902277278210	YNT DUUGOPIS ŻELOWE GSZT. PASTEL	SZT	2	6.40	12.80	23%	2.94	15.74	0.024	
28	A01	5902277278197	YNT DUUGOPIS ŻELOWE GSZT. METALLIC	SZT	2	6.40	12.80	23%	2.94	15.74	0.220	
29	A01	5902277265531	ZESZYTY A4 60# 70G PP	SZT	5	3.40	17.00	23%	3.91	20.91	1.500	
30	A01	5902277293962	N-ZESZYTY A4 60# M 70G ES KRAFT	SZT	5	3.15	15.75	23%	3.62	19.37	1.460	
31	A01	5902277175014	ZESZYTY A4 60# M 70G UV	SZT	5	3.14	15.70	23%	3.61	19.31	1.900	

KONIEC STRONY: 1

Rysunek 3.17: Tabela dok_int2_0 ze zbioru treningowego po binaryzacji



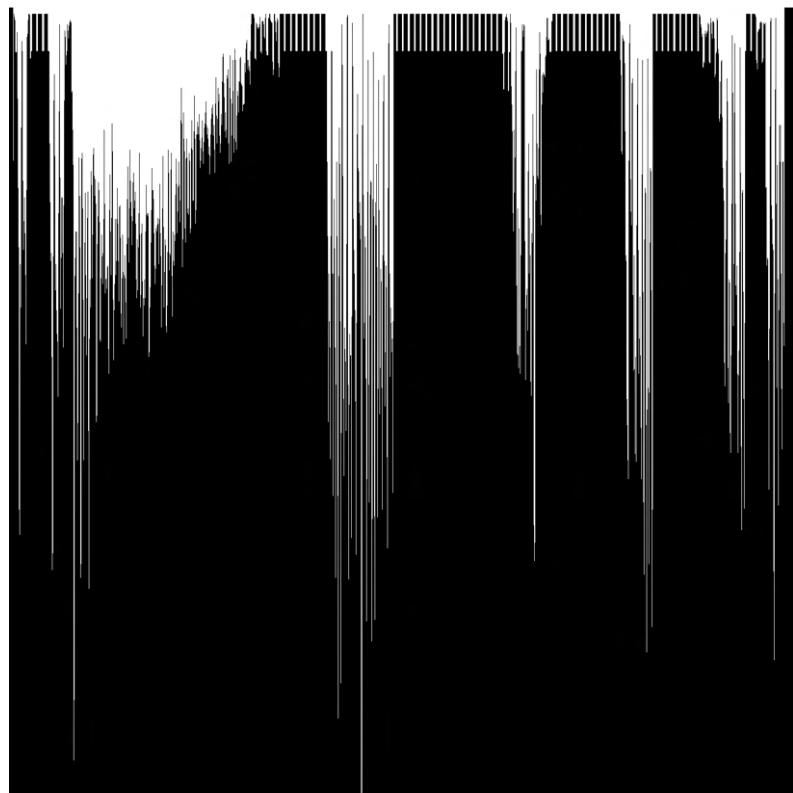
Rysunek 3.18: Projekcja pikseli tabeli z rys. 3.17 na oś poziomą



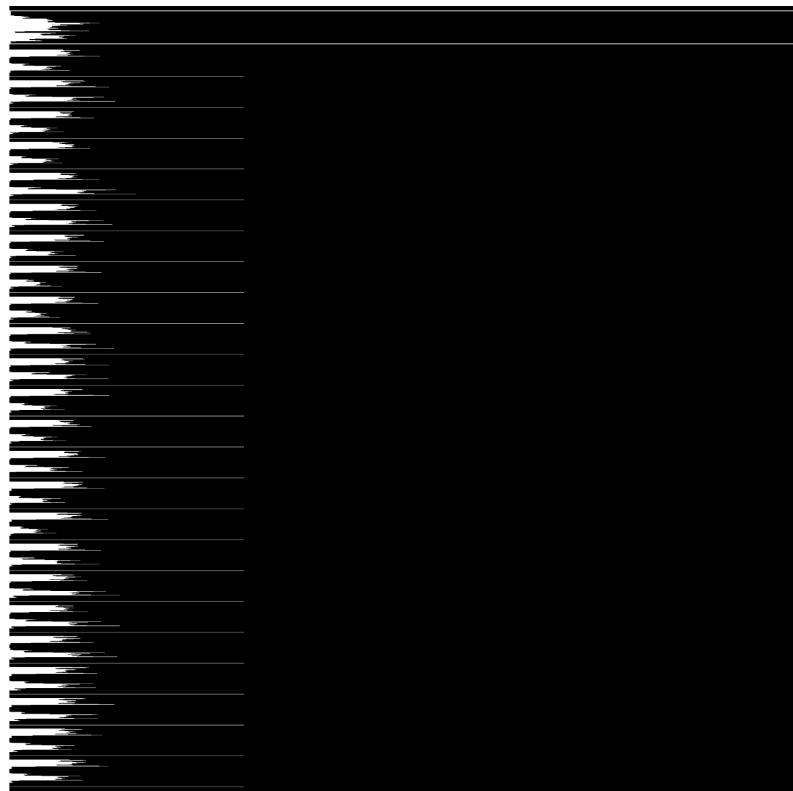
Rysunek 3.19: Projekcja pikseli tabeli z rys. 3.17 na oś pionową

Lp.	Kod Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
1	447 Blok milimetrowy A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515 Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4 Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1 Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540 Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,8100	38,10	23%
6	45 Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,1700	190,20	23%
7	422 Blok szkolny w linie A4-100	5905824900190	20 szt.	2,2500	45,00	23%
8	41 Blok tech. A3-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42 Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14 Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412 Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93 Brulion 300k. A4 tw opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99 Brulion 300k. A5 tw opr.	5905824005546	2 szt.	7,4200	14,84	23%
14	139 Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3000	58,60	23%
15	170 Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856 Karton A3-20 złoty	59058242020386	2 opak.	8,5500	17,10	23%
17	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594 Mapa Europy -szkol podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
19	595 Mapa Polski-szkol podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599 Mapa Swiata -szkol podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501 Pap.fluo.zółty samoprzyklejny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476 Papier srebrny samoprzyklejny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264 Papier (jk) A4-100 soledynowy	5905824020249	1 opak.	4,7500	4,75	23%
24	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.20: Tabela kr1_0 ze zbioru treningowego po binaryzacji



Rysunek 3.21: Projekcja pikseli tabeli z rys. 3.20 na oś poziomą



Rysunek 3.22: Projekcja pikseli tabeli z rys. 3.20 na oś pionową

Wykonane eksperymenty przyniosły ciekawe rezultaty. Z obrazów samej projekcji dało się wizualnie wywnioskować siatkę rozpatrywanych tabel. Co więcej, uznano, że do wnioskowania nie jest potrzebna dodatkowa analiza połączonych komponentów - wystarczyła odpowiednia filtracja i transformacja otrzymanych jednowymiarowych sygnałów. Oprócz tego na projekcjach dało się wyodrębnić rzuty obramowań, również tych przerywanych. Wyraźne różnice pomiędzy regionami tabel mogły wystarczyć do przeprowadzenia segmentacji w sposób progowy, bez wykorzystania złożonych narzędzi sztucznej inteligencji i uczenia maszynowego.

Otrzymane rezultaty i poczynione obserwacje uznano za wystarczająco przekonujące i podjęta została decyzja o kontynuowaniu prac z wykorzystaniem projekcji. Wybrane zostało podejście rokujące dobre rezultaty przy stosunkowo małej komplikacji realizacyjnej. Niewykluczone jednak, że podążenie inną ścieżką również mogło przynieść zamierzone efekty.

3.3.2 Własne modyfikacje algorytmów

Opracowanie złożonej, wielomodułowej metody osiągającej dobre i stabilne rezultaty wymagało przyjęcia pewnej strategii rozwijania jej. Jasnym było, że metodyka prac w dużym stopniu zadecyduje o efektach końcowych.

Prace rozpoczęto od skonstruowania szablonu metody (rys. 3.24), pokrywającego pełną ścieżkę przetwarzania od wejściowych plików pdf lub jpg po wyjściowy plik csv, wykorzystujący do segmentacji jedynie informacje o obramowaniach. Była to baza algorytmu, do której stopniowo wprowadzono dodatkowe tryby i poprawki wynikające z wyników testów.

W kolejnych etapach dodano moduły analizy tabel bez poziomych i pionowych obramowań i korygowano je poprzez zmianę metod filtracji, transformacji i wnioskowania w celu poprawienia rezultatów. Po wprowadzeniu każdej modyfikacji wyniki były wizualnie oceniane na wybranych 12 tabelach (każdej pochodzącej od innego dostawcy) ze zbioru treningowego, a w kluczowych momentach (np. po dodaniu nowego trybu przetwarzania lub doprowadzeniu algorytmu do bardzo dobrych wyników na wszystkich 12 podstawowych tabelach) sprawdzano rezultaty dla całego zbioru. Wynikające z obserwacji wnioski pozwalały decydować jaki powinien być kierunek modyfikacji.

W początkowych założeniach, metoda miała być poprawiana do momentu, w którym osiągnie bardzo dobre rezultaty na wszystkich tabelach treningowych bez wprowadzania innych parametrów dodatkowych poza domyślnymi. Po osiągnięciu zamierzonych efektów, miała zostać przeprowadzona ewaluacja i zestawienie z innymi metodami dostępnymi w publikacjach z wykorzystaniem własnego zbioru testowego. Analizy poczynione w trakcie rozwijania algorytmu pokazały, że faktury w zbiorze były mocno zróżnicowane i podjęte próby ujednolicenia parametrów dla każdej z nich skutkowały ogólnym pogorszeniem rezultatów.

Zdecydowano się pozostać przy rozwiązaniu parametryzowanym (na wzór biblioteki Came-

lot [2]), gdyż dawało ono możliwość dostrojenia i udanego zaaplikowania metody na szerszym zbiorze tabel (również tych różniących się stylem od przypadków w zbiorze). By rozwiążanie pozostało pozyteczne, przyjęto założenie, że faktury pochodzące z jednego źródła (wyróżniające się zbliżonym stylem) musiały być obsługiwane przez jeden zestaw parametrów. W ten sposób możliwe było zapisanie ustawień rozpoznawania tabeli dla każdego dostawcy i wtórne wykorzystywanie ich bez potrzeby manualnego dostrajania przez użytkownika.

3.3.3 Opracowane narzędzie

Prace rozpoczęto od opracowania aplikacji, która pozwoliła na przeprowadzenie badań, będąc jednocześnie środowiskiem rozwojowym docelowego algorytmu końcowego.

Zadaniem aplikacji było:

- wygodne wczytywanie dokumentów ze zbioru z ekstrakcją samych regionów tabeli z wykorzystaniem informacji z ground - truth
- przeprowadzenie pełnej ścieżki wnioskowania metody (rys. 3.24) na wczytanych obrazach
- dane możliwości szybkiej, wizualnej ewaluacji wyników cząstkowych jak i końcowych wywołanych przez wprowadzane modyfikacje
- przyjmowanie dodatkowych parametrów od użytkownika w celu dostrojenia metody do stylu faktury

Aplikacja w całości napisana została w języku Python. Do implementacji wykorzystano następujące biblioteki:

- pdf2image (konwersja plików pdf na zdjęcia)
- opencv-python (operacje morfologiczne na obrazie)
- pillow (obsługa plików obrazowych)
- numpy (większość potrzebnych obliczeń)
- scipy (metody filtracji sygnałów)
- pytesseract (moduł OCR)
- tkinter (interfejs użytkownika)

Aplikacja w finalnej wersji pozwala na wprowadzenie 11 parametrów należących do 3 grup (generalnej, przetwarzania pionowego, przetwarzania poziomego), od których uzależniony jest pełny pipeline przetwarzania obrazu:

- parametry generalne:
 - p_1 - próg binaryzacji z zakresu (0, 255)
 - p_2 - tryb przetwarzania pionowego (z obramowaniami, bez obramowań)
 - p_3 tryb przetwarzania poziomego (z obramowaniami, bez obramowań)
 - p_4 - tryb przetwarzania tła (czy korzystać z dodatkowych informacji o odcieniu tła)
 - p_5 - próg odcienia tła z zakresu (0, 255)
- parametry przetwarzania pionowego
 - p_6 - próg decyzyjny znormalizowany w osi pionowej z zakresu (0, 1)
 - p_7 - współczynnik wykładniczy transformacji potęgowej w osi pionowej (liczba naturalna)
 - p_8 - współczynnik długości filtra (liczba naturalna)
- parametry przetwarzania poziomego
 - p_9 - próg decyzyjny znormalizowany w osi poziomej z zakresu (0, 1)
 - p_{10} - współczynnik wykładniczy transformacji potęgowej w osi poziomej (liczba naturalna)
 - p_{11} - współczynnik długości filtra w osi poziomej (liczba naturalna)

Aplikacja pozwala na wyświetlenie obrazów przedstawiających 7 etapów działania algorytmu:

1. obraz po binaryzacji
2. surowa projekcja na oś pionową
3. surowa projekcja na oś pozioma
4. projekcja na oś pionową po filtracji i transformacjach
5. projekcja na oś poziomą po filtracji i transformacjach
6. otrzymana przez moduły wnioskujące krata tabeli
7. obraz końcowy po segmentacji

Domyślnie generowane są jedynie informacje o segmentacji. Wykorzystanie modułu OCR i wygenerowanie CSV było wywoływanie osobno by zaoszczędzić czas obliczeniowy przy ewaluacji wizualnej.

Tak przygotowana aplikacja pozwalała na wygodne modyfikowanie kluczowych modułów segmentacji i szybką ewaluację wprowadzanych zmian.

Lp	Nazwa towaru lub usługi	PKWiU	Ilość	Jedn. m.	Cena netto bez rabatu (zł)
1	Długopis S-FINE niebieski GWIAZDKI TO-059 TOMA dlk4900118		60	szt.	0,9
2	Długopis S-FINE automatyczny niebieski TO-069 TOMA dlk4540118		30	szt.	1,1
3	Naboje PARKER (5sztuk) niebieskie 195038 4 nak090055		6	szt.	7,6
4	Długopis FLEXI niebieski PENMATE TT7038 dlk4960094		60	szt.	0,8
5	Segregator A4/35mm 2 ringi czarny 14454 ESSELTE VIVIDA		2	szt.	7,7

Rysunek 3.23: Interfejs graficzny przygotowanej aplikacji

3.3.4 Zarys zaproponowanej metody rozpoznawania struktury

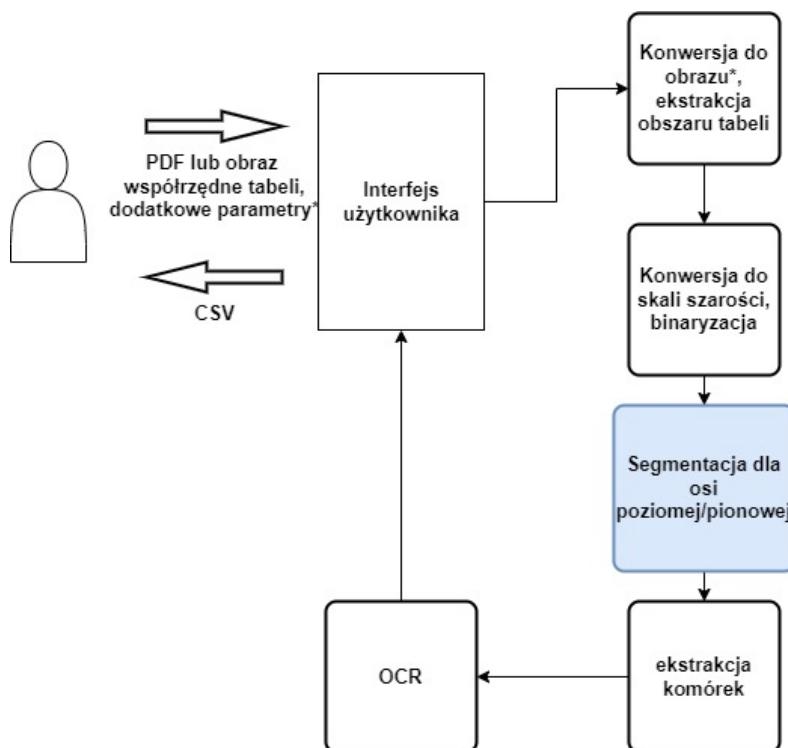
Ideę opracowanej metody oparto na algorytmie Zuyeva [36]. Wnioskowanie miało odbywać się poprzez analizę dolin projekcji profilowej obrazu dokumentu. Zmieniono jednak podstawowe założenie dotyczące projekcji, tzn. wykonywano rzutowanie na osie obrazu samych pikseli zamiast połączonych komponentów. Zdecydowano się na ten zabieg, kierując się intuicją, że analiza połączonych komponentów może być bardziej zawodna dla obrazów gorszej jakości (poprzez nieuchciane scalenia i podziały komponentów), co mogłoby powodować problemy przy przyszłych próbach uogólniania systemu.

Zmiana podejścia wymusiła wprowadzenie dodatkowych modyfikacji w procesie wstępnego przetwarzania sygnałów i wnioskowania. W całkowicie odmienny sposób ujęto również problem segmentacji wierszowej, co wynikało z faktu, że model zaproponowany przez Zuyeva nie miał szans poradzić sobie z tabelami z zaproponowanego zbioru.

Już na samym początku rozwiązano problem początkowego kroku binaryzacji obrazów. Przetestowano binaryzację progową dla kilku progów w standardowej skali (0, 255): 235, 200, 190. Dla progu 235 można było dostrzec pewne nieuchciane artefakty otaczające właściwe znaki, wynikające prawdopodobnie ze sposobu konwersji plików pdf do obrazów (biblioteka pdf2image).

Dla progu 200 problem artefaktów zniknął, ale w niektórych tabelach szary kolor tła powodował zakrycie właściwego tekstu. Ostatecznie próg 190 okazał się być idealnym rozwiązaniem dla wszystkich tabel ze zbioru treningowego.

Niezadowalające wyniki działania modułu OCR na pełnych obrazach tabel (rys. 3.12) rodziły pewne obawy co do możliwości praktycznego zastosowania wyników końcowego modelu. W trakcie pierwszych eksperymentów nie przywiązywano dużej wagi do rozpoznawanego tekstu. Jak się później okazało, po udoskonaleniu modułów segmentacji i przeprowadzeniu OCR na osobnych komórkach, wyniki działania *pytesseract* były ogólnie bardzo dobre (zdecydowana większość przeklamań znaków była sporadyczna, przewidywalna, łatwa do wykrycia i obsłużenia).



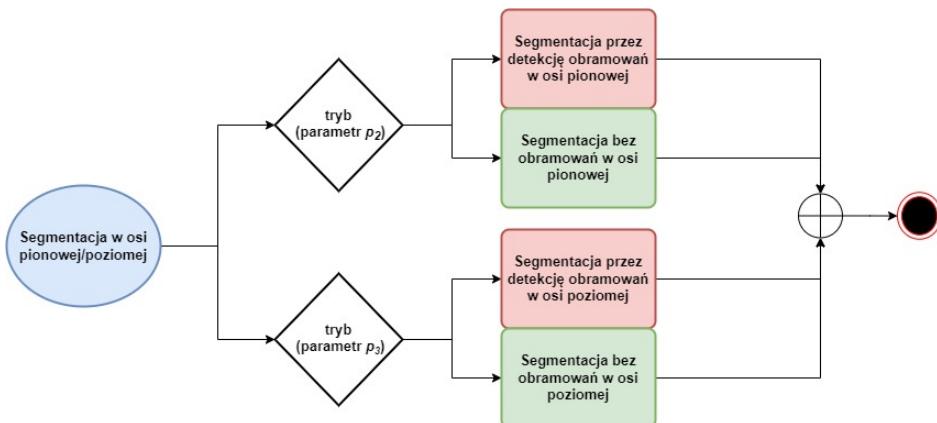
Rysunek 3.24: Ogólny schemat działania opracowanej metody

Kluczowymi i najbardziej wymagającymi były moduły segmentacji poziomej i pionowej obrazów tabeli. Segmentacja na każdej z osi przebiega w inny sposób. Dodatkowo, w każdym przypadku występują dwa tryby:

- wnioskowania na podstawie obramowań
- wnioskowania bez obramowań

Należy zaznaczyć, że w typowym podejściu, samowystarczalnym trybem jest wnioskowanie bez obramowań, który stosuje eliminację obramowań w pre - processingu. Tryb segmentacji z wykorzystaniem samych obramowań jest za to bardziej przewidywalny i działa bardzo dobrze dla

zdecydowanej większości przypadków z domyślnym zestawem parametrów (zakładając, że tabela jest w pełni obramowana).



Rysunek 3.25: Tryby segmentacji

3.3.5 Detekcja obramowań

Algorytm wykrywania obramowań oparty jest o elementy rozwiązań z literatury i własne spostrzeżenia.

Pionowe i poziome linie są dobrze widoczne na zdjęciach po zastosowaniu projekcji. W celu ich ekstrakcji wykorzystywane są równolegle 2 podejścia:

- ekstrakcja linii z wykorzystaniem morfologii
- prosty algorytm progowy zastosowany dla przetransformowanej projekcji

Podejście pierwsze oparte jest standardowych technikach operacji morfologicznych z wykorzystaniem podłużnego elementu strukturalnego i erozji [1],[2],[27].

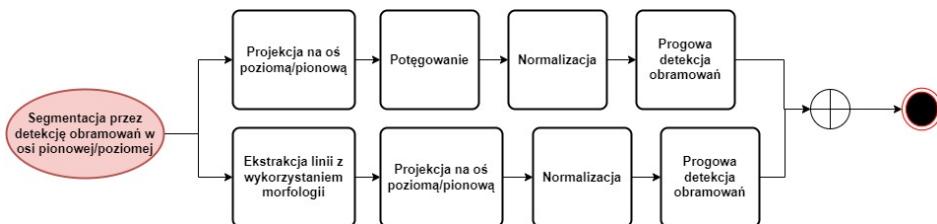
Wykonywane są kolejno:

1. erozja obrazu tabeli z elementem strukturalnym $1 \times p_6$ *szerokość obrazu (dla linii poziomych) lub $p_9 * \text{wysokość obrazu} \times 1$ (dla linii pionowych) pikseli
2. rzutowanie pozostawionych pikseli na oś poziomą (dla linii pionowych) lub pionową (dla linii poziomych)
3. progowa detekcja miejsc segmentacji

W tym wypadku, parametry p_6 i p_9 mają wpływ na to, jak długie linie będą rozpoznawane jako linie wyznaczające segmentacje - za duża wartość może spowodować ignorowanie krótszych linii (przykładowo w tabelkach podsumowujących), za mała może spowodować rozpoznawanie

tekstu jako linii separujących. Przyjęcie progu 0.1 działało idealnie w zdecydowanej większości przypadków.

Podejście to nie radzi sobie dobrze z liniami przerywanymi rysunki 3.27, 3.28. By uzupełnić brakujące części linii można stosować dylatację [1], jednak operacja ta, przy uzupełnianiu dużych luk może spowodować również zespolenie tekstu i błędne rozpoznawanie go jako linie.



Rysunek 3.26: Schemat działania detekcji obramowań

By poradzić sobie w tym przypadku, do algorytmu końcowego wprowadzono drugi pod - algorytm. Na obrazie projekcji 3.29 ludzkie oko jest w stanie wykryć linie separujące, mimo że na obrazie są one przerywane dużymi lukami. Postanowiono wykorzystać ten fakt i podjąć próbę wnioskowania na podstawie samego rzutowanego sygnału.

By ułatwić to zadanie i wyraźniej rozgraniczyć linie od pozostałych elementów obrazu zastosowano transformację potęgową. Funkcja potęgowa pozwala silnie wytłumić doliny sygnałów w odniesieniu do ich wartości charakterystycznych takich jak średnia czy wartość maksymalna.

Algorytm drugi składa się z następujących kroków:

1. surowe rzutowanie pikseli na oś poziomą (dla linii pionowych) lub pionową (dla linii poziomych)
2. normalizacja sygnału (do szerokości lub wysokości obrazu zależnie od osi przetwarzania)
3. potęgowanie z wykorzystaniem parametru p_7 lub p_{10} (zależnie od osi rzutowania)
4. progowa detekcja obramowań (z wykorzystaniem parametrów p_6 lub p_9)

Jak pokazują rysunki 3.30 i 3.31, opracowana metoda radzi sobie bardzo dobrze z przerywanymi liniami, ale wprowadzenie globalnego progu detekcji utrudnia znajdowanie krótszych linii, np. w tabelkach podsumowujących (rysunki 3.35 i 3.36).

Wyniki segmentacji otrzymane przez obie metody są sumowane, przez co niwelują fałszywie negatywne wyniki detekcji osiągane przez każdą z nich. Sumowane są również fałszywie pozytywne wyniki, ale testy na zbiorze treningowym nie wskazały żadnych przypadków, dla których wpłynęłyby to negatywnie na końcowe rezultaty (co świadczyło o dużej restrykcyjności obu metod).

Ogólne efekty były bardzo zadowalające. Należy jednak pamiętać że tabele ze zbioru były wygenerowane cyfrowo i wykrycie obramowań (nawet przerywanych) na dokumentach w tak dobrej jakości nie stanowi dużego wyzwania. Dla zdjęć lub skanów należałoby pamiętać o podjęciu dodatkowych kroków przetwarzania wstępnego, takich jak korekcja krzywizny, filtrowanie szumów i zastosowanie odpowiednio dobranej metody binaryzacji.

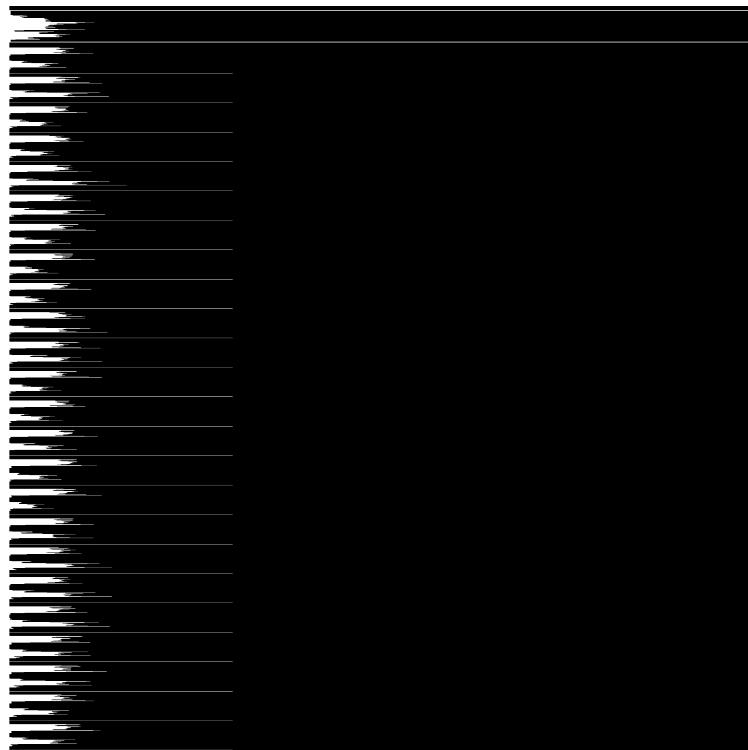
Lp.	Kod Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
1	447 Blok milimetrowy A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515 Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4 Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1 Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540 Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,8100	38,10	23%
6	45 Blok Superior techniczny kolorowy A4-25	590582410859	60 szt.	3,1700	190,20	23%
7	422 Blok szkolny w linie A4-100	5905824900190	20 szt.	2,2500	45,00	23%
8	41 Blok tech. A3-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42 Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14 Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412 Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93 Brulion 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99 Brulion 300k. A5 tw.opr.	5905824400546	2 szt.	7,4200	14,84	23%
14	139 Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3000	58,60	23%
15	170 Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856 Karton A3-20 złoty	5905824020386	2 opak.	8,5500	17,10	23%
17	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594 Mapa Europy -szkol.podkl.na biurko-intro	5905824400799	1 szt.	5,1300	5,15	5%
19	595 Mapa Polski -szkol.podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599 Mapa Świata -szkol.podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501 Pap.fluo.złoty samoprzylepny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476 Papier srebrny samoprzylepny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264 Papier (jk) A4-100 seledynowy	5905824020249	1 opak.	4,7300	4,75	23%
24	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.27: Tabela kr6_0 z przerywanymi liniami

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

L.p.	Kod Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
1	447 Blok millimetry A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515 Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4 Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1 Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540 Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,8100	38,10	23%
6	45 Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,1700	190,20	23%
7	422 Blok szkolny w linie A4-100	5905824900190	20 szt.	2,2500	45,00	23%
8	41 Blok tech. A4-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42 Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14 Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412 Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93 Brulion 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99 Brulion 300k. A5 tw.opr.	5905824400546	2 szt.	7,4200	14,84	23%
14	139 Kalendarz VITO A4 chamois	5905824202669	2 szt.	29,3000	58,60	23%
15	170 Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856 Karton A3-20 złoty	5905824020386	2 opak.	8,5500	17,10	23%
17	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594 Mapa Europy -szkol.podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
19	595 Mapa Polski -szkol.podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599 Mapa Świata -szkol.podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501 Pap.fluo.złoty samoprzyklejny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476 Papier srebrny samoprzyklejny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264 Papier (jk) A4-100 szledynowy	5905824020249	1 opak.	4,7500	4,75	23%
24	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.28: Segmentacja tabeli z rysunku 3.27 w osi pionowej metody opartej o morfologię



Rysunek 3.29: Projekcja tabeli 3.27 na oś pionową

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA



Rysunek 3.30: Projekcja tabeli 3.27 na oś pionową po transformacji (w stosunku do średniej)

Lp.	Kod Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
1	447 Blok milimetrowy A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515 Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4 Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1 Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540 Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,8100	38,10	23%
6	45 Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,1700	190,20	23%
7	422 Blok szkolny w linie A4-100	5905824900190	20 szt.	2,2500	45,00	23%
8	41 Blok tech. A3-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42 Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14 Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412 Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93 Brulion 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99 Brulion 300k. A5 tw.opr.	5905824400546	2 szt.	7,4200	14,84	23%
14	139 Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3000	58,60	23%
15	170 Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856 Karton A3-20 złoty	5905824020386	2 opak.	8,5500	17,10	23%
17	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594 Mapa Europy -szkol.podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
19	595 Mapa Polski -szkol.podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599 Mapa Świata -szkol podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501 Pap.fluo żółty samoprzylepny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476 Papier srebrny samoprzylepny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264 Papier (jk) A4-100 seledynowy	5905824020249	1 opak.	4,7500	4,75	23%
24	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.31: Wyniki segmentacji tabeli 3.27 w osi pionowej przez metodę wykorzystującą projekcję

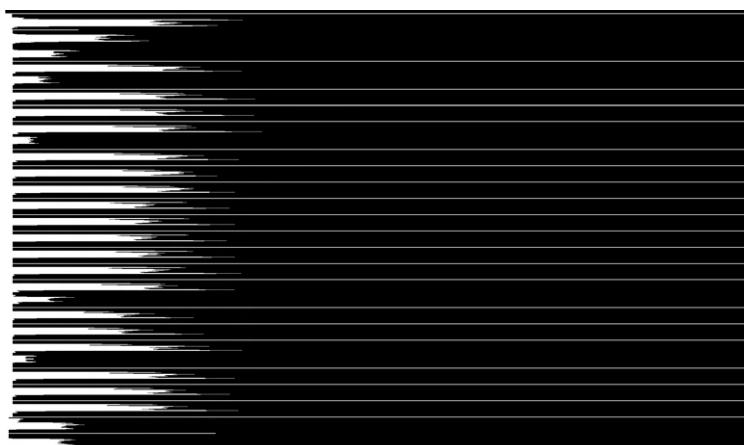
ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

LP	INDEKS	NAZWA - OPIS TOWARU / USŁUGI	JM	ILOŚĆ	CENA NETTO	WARTOŚĆ POZYCJI BEZ %	PODAZEK VAT %	WARTOŚĆ POZYCJI Z PODATKIEM
1	obk1950089	Koszulki kryształiczne BANTEX A4 45mic., w kartonie (100szt) 100550096	op	20.000	9.65	193.00	23 44.39	237.39
2	tok0110070D	Torebki strunowe (100sztuk) 230x320mm 35mic DOTTs	op	10.000	12.24	122.40	23 28.15	150.55
3	tok0110070D	Torebki strunowe (100sztuk) 230x320mm 35mic DOTTs	op	10.000	12.24	122.40	23 28.15	150.55
4	obk5002089	Koszulki groszkowe A4 35mic. BANTEX BUDGET (100szt) 400105682	op	20.000	3.75	75.00	23 17.25	92.25
5	mak6080344	Marker permanentny PILOT JUMBO czarny PISC-66B	szt.	4.000	12.64	50.56	23 11.63	62.19
6	mak2760082	Marker kredowy UNI PWE-8K biały 8mm UNPWEBK/6B1	szt.	1.000	9.44	9.44	23 2.17	11.61
7	mak2760082	Marker kredowy UNI PWE-8K biały 8mm UNPWEBK/6B1	szt.	3.000	9.44	28.32	23 6.51	34.83
8	zak0120022	Zakrzesacz TOPSTAR turkus 364-35 STAEDTLER	szt.	1.000	3.39	3.39	23 0.78	4.17
9	zak0120022	Zakrzesacz TOPSTAR turkus 364-35 STAEDTLER	szt.	2.000	3.39	6.78	23 1.56	8.34
10	zak0100022	Zakrzesacz TOPSTAR fiolet 364-6 STAEDTLER	szt.	3.000	3.39	10.17	23 2.34	12.51
11	zak0070022	Zakrzesacz TOPSTAR niebieski 364-3 STAEDTLER	szt.	3.000	3.39	10.17	23 2.34	12.51
12	zak0060022	Zakrzesacz TOPSTAR czerwony 364-2 STAEDTLER	szt.	2.000	3.39	6.78	23 1.56	8.34
13	skk3220239D	Skoroszyt zawieszany PP DOTTs (20) niebieski wzmacniony polipropy	op	4.000	7.57	30.28	23 6.96	37.24
14	ok 0210153	Okładka na dokumenty brokuc.D1	szt.	2.000	2.78	5.56	23 1.28	6.84
15	ok 0210153	Okładka na dokumenty brokuc.D1	szt.	8.000	2.78	22.24	23 5.12	27.36
16	ok 0054019	Okładka na dokumenty z dowodem rej. czarne OD-20-05 BIURFOL	szt.	2.000	2.35	4.70	23 1.06	5.78
17	ok 0124019	Okładka na dokumenty mini mro OD-25-03 BIURFOL (X)	szt.	5.000	3.19	15.95	23 3.67	19.62
18	ok 0187019	Okładka na dokumenty mini sky KOD-03-06 BIURFOL	szt.	3.000	2.98	8.94	23 2.06	11.00
19	ok 0183019	Okładka na dokumenty mini grass KOD-03-02 BIURFOL	szt.	2.000	2.98	5.96	23 1.37	7.33
(1253)			Razem		732.04	x 168.37	900.41	
			W tym		732.04	23 168.37	900.41	

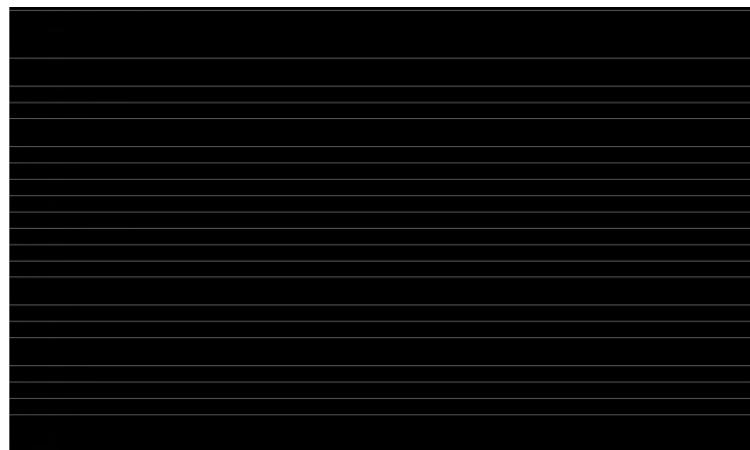
Rysunek 3.32: Tabela gd7e9pWKy0yL_FV35231H2A2021_0 z tabelką podsumowującą

LP	INDEKS	NAZWA - OPIS TOWARU / USŁUGI	JM	ILOŚĆ	CENA NETTO	WARTOŚĆ POZYCJI BEZ %	PODAZEK VAT %	WARTOŚĆ POZYCJI Z PODATKIEM
1	obk1950089	Koszulki kryształiczne BANTEX A4 45mic., w kartonie (100szt) 100550096	op	20.000	9.65	193.00	23 44.39	237.39
2	tok0110070D	Torebki strunowe (100sztuk) 230x320mm 35mic DOTTs	op	10.000	12.24	122.40	23 28.15	150.55
3	tok0110070D	Torebki strunowe (100sztuk) 230x320mm 35mic DOTTs	op	10.000	12.24	122.40	23 28.15	150.55
4	obk5002089	Koszulki groszkowe A4 35mic. BANTEX BUDGET (100szt) 400105682	op	20.000	3.75	75.00	23 17.25	92.25
5	mak6080344	Marker permanentny PILOT JUMBO czarny PISC-66B	szt.	4.000	12.64	50.56	23 11.63	62.19
6	mak2760082	Marker kredowy UNI PWE-8K biały 8mm UNPWEBK/6B1	szt.	1.000	9.44	9.44	23 2.17	11.61
7	mak2760082	Marker kredowy UNI PWE-8K biały 8mm UNPWEBK/6B1	szt.	3.000	9.44	28.32	23 6.51	34.83
8	zak0120022	Zakrzesacz TOPSTAR turkus 364-35 STAEDTLER	szt.	1.000	3.39	3.39	23 0.78	4.17
9	zak0120022	Zakrzesacz TOPSTAR turkus 364-35 STAEDTLER	szt.	2.000	3.39	6.78	23 1.56	8.34
10	zak0100022	Zakrzesacz TOPSTAR fiolet 364-6 STAEDTLER	szt.	3.000	3.39	10.17	23 2.34	12.51
11	zak0070022	Zakrzesacz TOPSTAR niebieski 364-3 STAEDTLER	szt.	3.000	3.39	10.17	23 2.34	12.51
12	zak0060022	Zakrzesacz TOPSTAR czerwony 364-2 STAEDTLER	szt.	2.000	3.39	6.78	23 1.56	8.34
13	skk3220239D	Skoroszyt zawieszany PP DOTTs (20) niebieski wzmacniony polipropy	op	4.000	7.57	30.28	23 6.96	37.24
14	ok 0210153	Okładka na dokumenty brokuc.D1	szt.	2.000	2.78	5.56	23 1.28	6.84
15	ok 0210153	Okładka na dokumenty brokuc.D1	szt.	8.000	2.78	22.24	23 5.12	27.36
16	ok 0054019	Okładka na dokumenty z dowodem rej. czarne OD-20-05 BIURFOL	szt.	2.000	2.35	4.70	23 1.06	5.78
17	ok 0124019	Okładka na dokumenty mini mro OD-25-03 BIURFOL (X)	szt.	5.000	3.19	15.95	23 3.67	19.62
18	ok 0187019	Okładka na dokumenty mini sky KOD-03-06 BIURFOL	szt.	3.000	2.98	8.94	23 2.06	11.00
19	ok 0183019	Okładka na dokumenty mini grass KOD-03-02 BIURFOL	szt.	2.000	2.98	5.96	23 1.37	7.33
(1253)			Razem		732.04	x 168.37	900.41	
			W tym		732.04	23 168.37	900.41	

Rysunek 3.33: Segmentacja tabeli z rysunku 3.32 w osi pionowej metody opartej o morfologię



Rysunek 3.34: Projekcja tabeli 3.32 na oś pionową



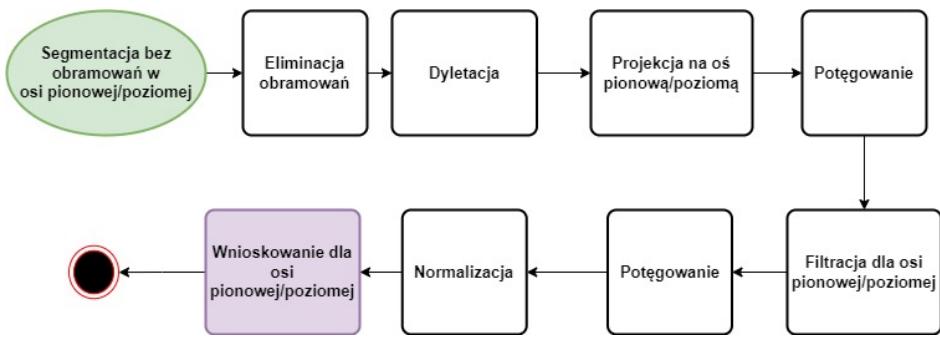
Rysunek 3.35: Projekcja tabeli 3.32 na oś pionową po transformacji (w stosunku do średniej)

Lp	INDEKS	NAZWA - OPIS TOWARU / USŁUGI	JM	ILOŚĆ	CENA NETTO	WARTOŚĆ POZYCJI BEZ PODATKU	PODAZEK VAT %	WARTOŚĆ POZYCJI Z PODATKIEM
1	obk1950089	Koszulki kryształiczne BANTEX A4 45mic., w kartonie (100sztuk) 100550096	op	20.000	9.65	193.00	23	44.39
2	tok01100070D	Torebki strunowe (100sztuk) 230x320mm 35mic DOTTs	op	10.000	12.24	122.40	23	28.15
3	tok01100070D	Torebki strunowe (100sztuk) 230x320mm 35mic DOTTs	op	10.000	12.24	122.40	23	28.15
4	obk5020089	Koszulki groszkowe A4 35mic. BANTEX BUDGET (100szt) 400105682	op	20.000	3.75	75.00	23	17.25
5	mak6080344	Marker permanentny PILOT JUMBO czarny PISC-66B	szt.	4.000	12.64	50.56	23	11.63
6	mak2760082	Marker kredowy UNI PWE-8K biały 8mm UNPWE8K/8B	szt.	1.000	9.44	9.44	23	2.17
7	mak2760082	Marker kredowy UNI PWE-8K biały 8mm UNPWE8K/8B	szt.	3.000	9.44	28.32	23	6.51
8	zak0120022	Zakreślacz TOPSTAR turkus 364-35 STAEDTLER	szt.	1.000	3.39	3.39	23	0.78
9	zak0120022	Zakreślacz TOPSTAR turkus 364-35 STAEDTLER	szt.	2.000	3.39	6.78	23	1.56
10	zak0100022	Zakreślacz TOPSTAR fiolet 364-6 STAEDTLER	szt.	3.000	3.39	10.17	23	2.34
11	zak0070022	Zakreślacz TOPSTAR niebieski 364-3 STAEDTLER	szt.	3.000	3.39	10.17	23	2.34
12	zak0060022	Zakreślacz TOPSTAR czerwony 364-2 STAEDTLER	szt.	2.000	3.39	6.78	23	1.56
13	skk3220239D	Skoroszyt zawieszany PP DOTTs (20) niebieski wzmacniony poliprop.	op	4.000	7.57	30.28	23	6.96
(1253)						Razem		37.24
						732.04	x	168.37
						732.04		900.41
						732.04	23	168.37
								900.41

Rysunek 3.36: Wyniki segmentacji tabeli 3.32 w osi pionowej przez metodę wykorzystującą potęgowanie i projekcję

3.3.6 Wnioskowanie bez korzystania z obramowań

Ogólny zarys algorytmu jest bardzo zbliżony zarówno dla osi poziomej jak i pionowej. Pipeline wstępnego przetwarzania obu sygnałów jest identyczny, w praktyce różni się tylko wartościami parametrów wejściowych. Znaczące różnice algorytmów pojawiają się dopiero w modułach wnioskujących.



Rysunek 3.37: Schemat działania segmentacji bez obramowań

Pierwszą, podstawową czynnością jest usunięcie wszystkich obramowań z obrazu wykrytych metodą opisaną w rozdziale 3.3.5. Ma ono na celu zapobieganie myleniu obramowań z zawartością tabeli, gdyż to projekcja tekstu jest podstawowym źródłem informacji, na podstawie których następuje wnioskowanie. Lokalizacje obramowań są zapamiętywane i ponownie wykorzystywane pod koniec procesu segmentacji.

W kolejnym kroku wykonywana jest dylatacja na całym obrazie z wykorzystaniem elementu strukturyzującego rozmiarów 10×10 pixeli, mająca na celu uwydatnienie znaków. Operacja ta jest szczególnie potrzebna przy bardzo cienkich czcionkach. Rdzeń dylatacji został dobrany eksperymentalnie - testowane były kwadraty o rozmiarach 5×5 , 10×10 i 15×15 pikseli, jednak wizualna ocena wyników pokazała, że rdzeń 10×10 najlepiej współpracuje z dalszym przetwarzaniem zastosowanym w algorytmie dla większości tabel ze zbioru treningowego.

Po dokonaniu projekcji obrazu na odpowiednią oś, otrzymany sygnał jest poddany wstępemu potęgowaniu z wykładnikiem p_7 lub p_{10} , którego skutkiem jest redukcja szumów (niewykryte pozostałości obramowań, kawałki tekstu spoza obszaru tabeli) i innych niepożądanych elementów (skutki uboczne dylatacji, komórki nachodzące na sąsiadów, komórki zespolone). Krok ten jest szczególnie istotny z punktu widzenia następującej po nim metody filtrowania sygnału.

Surowy rzut obrazu cechuje się częstymi i dużymi fluktuacjami, przez co wnioskowanie na jego podstawie nie jest możliwe bez odpowiedniej filtracji. Rezultatem filtracji powinny być możliwe wypukłe wzniesienia odpowiadające blokom tekstu w rzutowaniu na oś poziomą i strome, wypukłe szczyty odpowiadające liniom tekstu w rzutowaniu na oś pionową. Wypróbowano kilka rodzajów filtrów górnoprzepustowych:

- filtr średniej ruchomej
- filtr IIR kaskadowy o symetrycznych warunkach brzegowych (`scipy.signal.symiirorder2`)
- filtr Chebysheva II rodzaju (`scipy.signal.cheby2`)

Pomimo manipulowania charakterystykami filtrów, żadne z wymienionych podejść nie przyniosło

zamierzonego rezultatu. Sukces przyniosło dopiero zastosowanie filtracji dwuetapowej z wykorzystaniem regułowej modyfikacji filtra ruchomego maksimum.

W pierwszej iteracji, stosowany jest filtr wycentrowanego ruchomego maksimum z oknem długości p_8 lub p_{11} , z zastrzeżeniem, że filtrowaniu poddawane są jedynie próbki sygnału przekraczające ustaloną, małą wartość (przyjęto 10% odchylenia standardowego sygnału wejściowego). W ten sposób, bloki tekstu nie są rozciągane na miejsca, w których nie wykryto nic poza domniemaną ciszą. Powodzenie tego kroku w dużej mierze oparte jest o dobre dostrojenie parametru potęgowania wykonywanego w poprzednim kroku, gdyż to ono decyduje, jak dobrze rozgraniczone są fragmenty z tekstem od fragmentów pustych.

Naturalnie, może zdarzyć się, że lokalne minima w rzucie tekstu nie zostaną objęte przez filtr, co poskutkuje powstaniem krótkich, niepożądanych przerw w przebiegu sygnału. By temu zapobiec sygnał jest filtrowany ponownie filtrem wycentrowanego ruchomego maksimum, tym razem bezwarunkowo i z 3 razy krótszym oknem filtra. W ten sposób luki są wypełniane, a zakres rzutu tekstu ulega jedynie niewielkiej zmianie, najczęściej nieistotnej z punktu widzenia dalszego przetwarzania.

Przefiltrowany sygnał poddawany jest normalizacji. W tym wypadku jest to dzielenie przez średnią. Obserwacje pokazały, że porównywanie wartości sygnału w zestawieniu z jego średnią wartością pozwala bardzo wyraźnie rozgraniczyć bloki tekstu i zwizualizować w odpowiednim przybliżeniu jego przebieg już po wyeliminowaniu niepożądanych wahań.

Sygnał wynikowy otrzymany tą metodą może zostać poddany procedurze wnioskowania. Jak jednak wspominano, sygnały w osi poziomej i pionowej znaczaco różnią się od siebie, dlatego metody wnioskowania również muszą być osobno przygotowane dla obu przypadków.

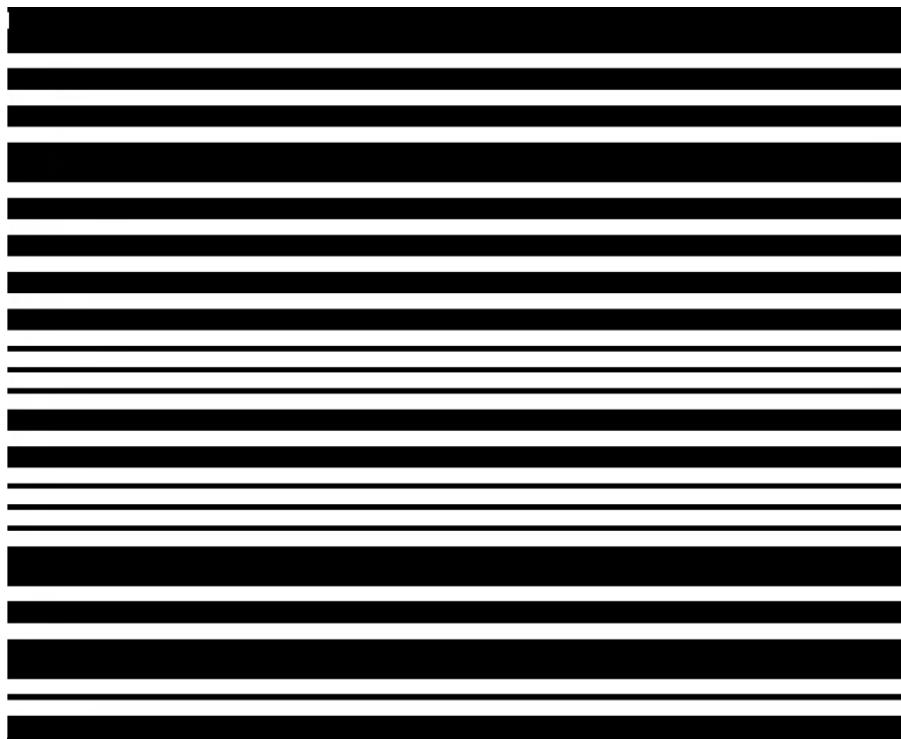
ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

Lp	Symbol	EAN	Symbol obcy	Opis	JM	Ilość	Cena PLN	Netto PLN	Vat		
									%	PLN	
1	HA 3720 2030-M10	5905130007736		Blok malarski Młody Artysta, A4, 10 ark, 200g, Happy Color	szt	5	3,09	15,45	23%	3,55	19,00
2	HA AKPB1471-3	6933615145395		Długopis usuwalny "buźki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	2,63	31,56	23%	7,26	38,82
3	HA AKPB1474-3	6941025162721		Długopis usuwalny, seria Space/Story, 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	2,80	33,60	23%	7,73	41,33
4	HA AKPA6571-3	6941025110098		Długopis usuwalny "kaczki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	2,63	31,56	23%	7,26	38,82
5	HA AKPB4471-3	6941025125399		Długopis usuwalny "uszaki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	36	2,63	94,68	23%	21,78	116,46
6	HA AKPB7371-3	6941025125429		Długopis usuwalny "zaryfy", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	24	2,63	63,12	23%	14,52	77,64
7	FO 77894	4001868073010		Drutki pluszowe MIX zielony, 10 szt, dł. 50cm, śr. 8mm, Folia	opk	2	3,29	6,58	23%	1,51	8,09
8	LA L300-95	4024526003501		Dziurkacz L300, 25 kartek, czarny, Laco	szt	1	15,52	15,52	23%	3,57	19,09
9	LA L300-25	4024526003525		Dziurkacz L300, 25 kartek, czerwony, Laco	szt	1	15,52	15,52	23%	3,57	19,09
10	HA 7370 0075-7	5905130006333		Farba akrylowa 75ml, brązowy, Happy Color	szt	1	3,70	3,70	23%	0,85	4,55
11	HA 7370 0075-75	5905130007262		Farba akrylowa 75ml, ciemnobrązowy, Happy Color	szt	1	3,70	3,70	23%	0,85	4,55
12	S 317-9	4007817310748		Folipops Lumocolor, M, wodoodporne, czarne, Staedtler	szt	10	4,28	42,80	23%	9,84	52,64
13	S 250 07-HB	4007817213889		Graffiti 0,7 mm, na papier, HB, Staedtler	szt	12	2,19	26,28	23%	6,04	32,32
14	S 250 05-2B	4007817213520		Graffiti 0,5 mm, na papier, 2B, Staedtler	szt	12	2,19	26,28	23%	6,04	32,32
15	S 250 05-HB	4007817213582		Graffiti 0,5 mm, na papier, HB, Staedtler	szt	12	2,19	26,28	23%	6,04	32,32
16	LA H400-35	4024526000029		Zszywacz metalowy H400, 30 kartek, głębokość zszywania 64 mm, niebieski, Laco	szt	1	12,37	12,37	23%	2,85	15,22
17	S 144 NC12 SET6	4007817609019		Kredki szkolne Noris Club, 12 kolorów + ołówek + gumka, Staedtler	szt	2	6,92	13,84	23%	3,18	17,02
18	HA AKR67K35-3	6937168812296		Wiklady do długopisu usuwalnego, Standard A, 0,5mm, niebieski, 3 szt. w etui, Happy Color	szt	30	4,91	147,30	23%	33,88	181,18
19	S 120-HB	4007817104620		Ołówek Noris, HB, nr 2, Staedtler	szt	12	1,34	16,08	23%	3,70	19,78
20	S 350-2	4007817321485		Marker Lumocolor, wodoodporny, gruby, ścięta końcówka, czerwony, Staedtler	szt	10	2,27	22,70	23%	5,22	27,92

Rysunek 3.38: Tabela 2018_FA_MG01_24115_0 pionowo obramowana ze zbioru treningowego



Rysunek 3.39: Surowa projekcja tabeli z rysunku 3.38 na oś pionową

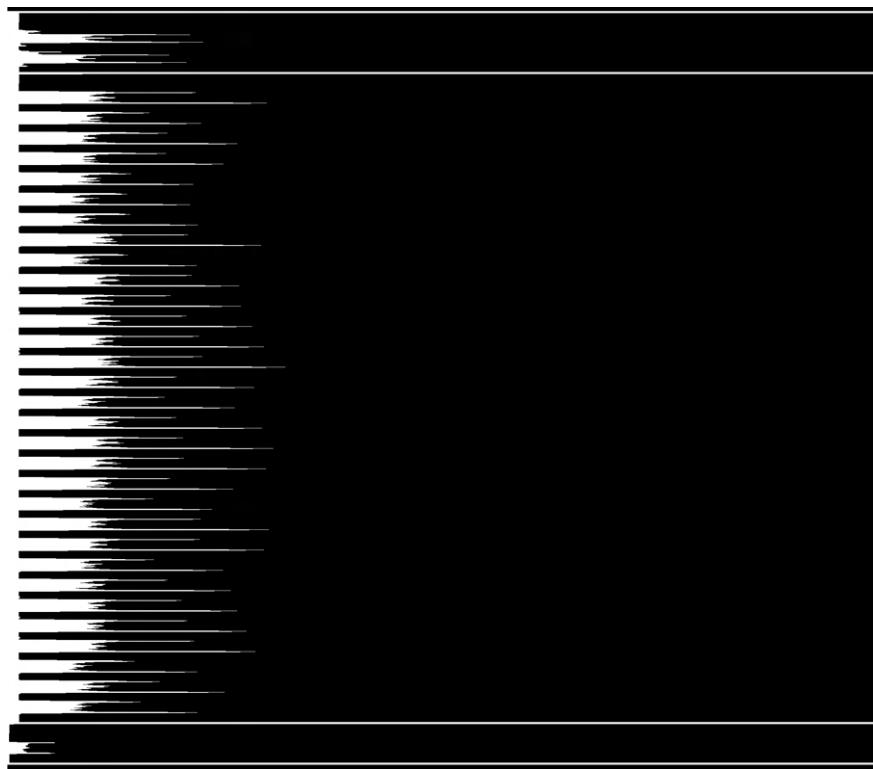


Rysunek 3.40: Sygnał z rysunku 3.39 przetransformowany i znormalizowany w zakresie wartości (0,1)

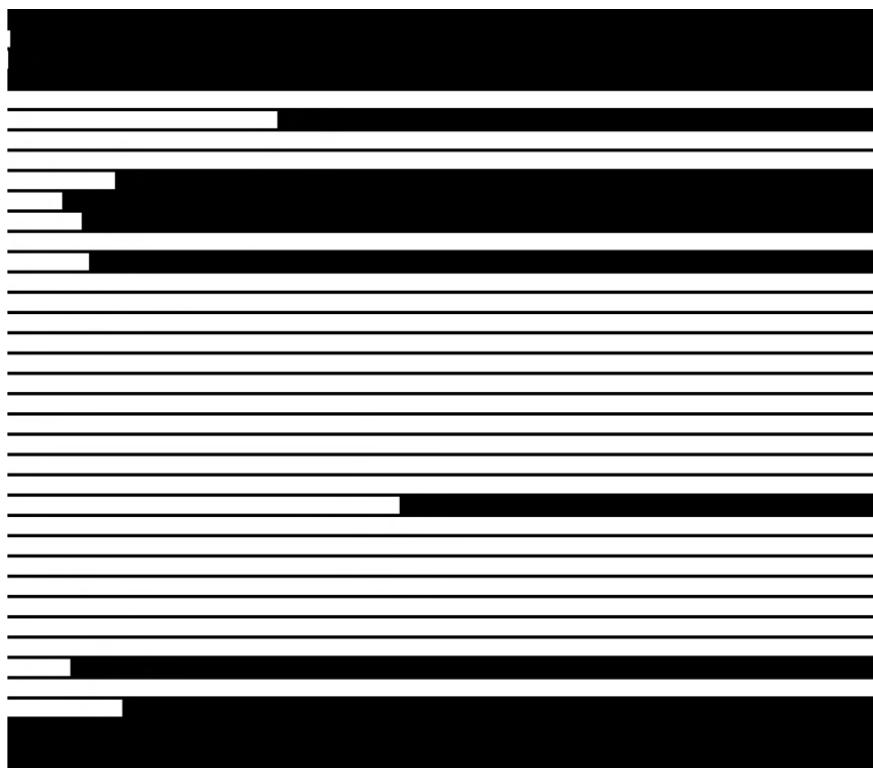
lp.	mag	kod kreskowy	nazwa towaru/usługi	kod CN /PKWIU	j.m	ilość	cena netto	wartość netto	% VAT	kwota vat	wartość brutto	masa (kg.)
1	A01	5902277274458	BBI LINIJKA FLEX 15CM B&B KIDS PASTEL		SZT	20	0.83	16.60	23%	3.82	20.42	0.300
2	A01	5902277060006	BLOK RYSUNKOWY A4 20 80G		SZT	40	0.84	33.60	23%	7.73	41.33	5.440
3	A01	5902277070012	BLOK TECHNICZNY A3 10 170G		SZT	120	1.95	234.00	23%	53.82	287.82	31.200
4	A01	5902277070005	BLOK TECHNICZNY A4 10 170G		SZT	120	0.97	116.40	23%	26.77	143.17	15.600
5	A01	5902277170828	BRULION A4 96# M 70G		SZT	10	4.46	44.60	23%	10.26	54.86	6.100
6	A01	5902277171726	BRULION A4 96# M 70G		SZT	5	4.46	22.30	23%	5.13	27.43	3.050
7	A01	5902277170804	BRULION A5 96# M 70G		SZT	10	2.35	23.50	23%	5.41	28.91	2.920
8	A01	5902277294241	BRULION A5 96# M 70G METALLIC SATIN GOLD		SZT	5	2.96	14.80	23%	3.40	18.20	1.460
9	A01	5902277170606	BRULION A6 96# M 70G		SZT	10	2.24	22.40	23%	5.15	27.55	1.450
10	A01	5902277179760	DIENNIK KORESPONDENCYJNY A4 96 70G		SZT	5	7.36	36.80	23%	8.46	45.26	2.700
11	A01	5902277295118	IBI OŁÓWEK B&B KIDS PASTEL		SZT	72	0.93	66.96	23%	15.40	82.36	0.432
12	A01	5902277295262	IBI OŁÓWEK ZE ZWIERZAKIEM B&B		SZT	36	1.90	68.40	23%	15.73	84.13	0.504
13	A01	5902277276551	IND.ZAKŁ.INDeksujące FUNKY 200 12X45MM		SZT	3	2.72	8.16	23%	1.88	10.04	0.033
14	A01	5902277274625	KOLOROWANKA Z NAKL.A4 16 DUŻO DO KOLOR.	58.11.13.0	SZT	10	2.06	20.60	5%	1.03	21.63	1.130
15	A01	5902277207418	KOLIZ.PP A5 100# SPIR.PO KR.BOKU Z G&B		SZT	5	5.81	29.05	23%	6.68	35.73	1.145
16	A01	5902277171337	KOŁOZESZYT A5 160# M 70G Z PERF.		SZT	5	4.23	21.15	23%	4.86	26.01	1.750
17	A01	5902277174437	KOSTKA PAP.8.5x8.5x3.5CM BIAŁA KLEJONA		SZT	30	1.25	37.50	23%	8.63	46.13	8.190
18	A01	5902277170569	KOSTKA PAP.8.5x8.5x3.5CM KOLOR KLEJONA		SZT	60	1.63	97.80	23%	22.49	120.29	12.180
19	A01	5902277170804	N-KOSTKA PAP.8.5x8.5x3.5CM KOLOR NIEKLEJ		SZT	10	1.63	16.30	23%	3.75	20.05	2.200
20	A01	5902277244475	NSR KLEJ BROKATOWY NOSTER 6ml a'5		SZT	3	1.49	4.47	23%	1.03	5.50	0.195
21	A01	5902277100009	PAPIER KOLOROWY A5 10 115G		SZT	10	0.67	6.70	23%	1.54	8.24	0.500
22	A01	5902277275035	TDS DŁUGOPIS TODAYS JET LINE BLACK a'10		SZT	10	0.99	9.90	23%	2.28	12.18	0.100
23	A01	5902277275042	TDS DŁUGOPIS TODAYS JET LINE BLUE a'10		SZT	10	0.99	9.90	23%	2.28	12.18	0.100
24	A01	5902277170118	WLAD DO SEG A4 50# KOL.M 70G		SZT	5	2.66	13.30	23%	3.06	16.36	1.175
25	A01	5902277244543	YNJ BROKAT SYPKI 7G A'6 PEARL		SZT	3	6.39	19.17	23%	4.41	23.58	0.315
26	A01	5902277278203	YNT DŁUGOPIS ŻELOWE 6SZT. NEON		SZT	2	6.40	12.80	23%	2.94	15.74	0.024
27	A01	5902277278210	YNT DŁUGOPIS ŻELOWE 6SZT. PASTEL		SZT	2	6.40	12.80	23%	2.94	15.74	0.024
28	A01	5902277278197	YNT DŁUGOPIS ŻELOWE 6SZT. METALLIC		SZT	2	6.40	12.80	23%	2.94	15.74	0.220
29	A01	5902277265531	ZESZYT A4 60# 70G PP		SZT	5	3.40	17.00	23%	3.91	20.91	1.500
30	A01	5902277293862	N-ZESZYT A4 60# M 70G HS KRAFT		SZT	5	3.15	15.75	23%	3.62	19.37	1.460
31	A01	5902277175014	ZESZYT A4 80# M 70G UV		SZT	5	3.14	15.70	23%	3.61	19.31	1.900

KONIEC STRONY: 1

Rysunek 3.41: Tabela dok_interdruk2_0 pionowo obramowana ze zbioru treningowego



Rysunek 3.42: Surowa projekcja tabeli z rysunku 3.41 na oś pionową

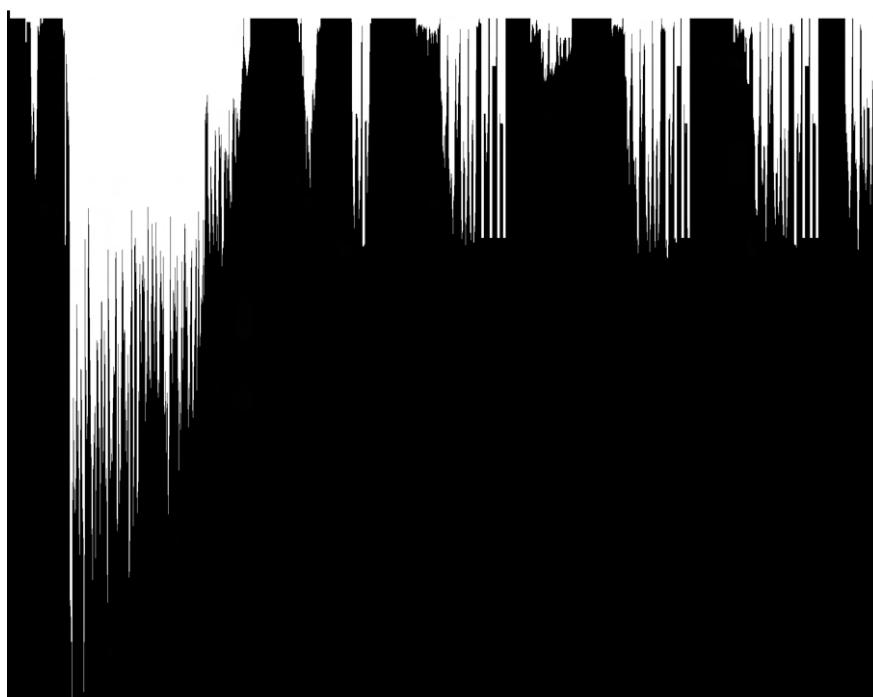


Rysunek 3.43: Sygnał z rysunku 3.42 przetransformowany i znormalizowany w zakresie wartości $(0,1)$

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamentowa biel 20 szt./op. 220g/m² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%
6	Folia laminacyjna PE arkusz A4 125 mic standard błysk antystatyczna Kod Indeksu: 320412	1	OP	29,50 PLN		29,50 PLN	29,50 PLN	23%
7	HEYKKA Zakreślacz klasyczny Linea jasnozielony, 10 szt./opk. Kod Indeksu: 611109	1	OP	21,00 PLN	35%	13,65 PLN	13,65 PLN	23%
8	HEYKKA Zakreślacz klasyczny Linea pastelowy żółty 10 szt./opk. Kod Indeksu: 611159	1	OP	21,00 PLN	35%	13,65 PLN	13,65 PLN	23%

Rysunek 3.44: Tabela FV210310097 poziomo obramowana ze zbioru treningowego



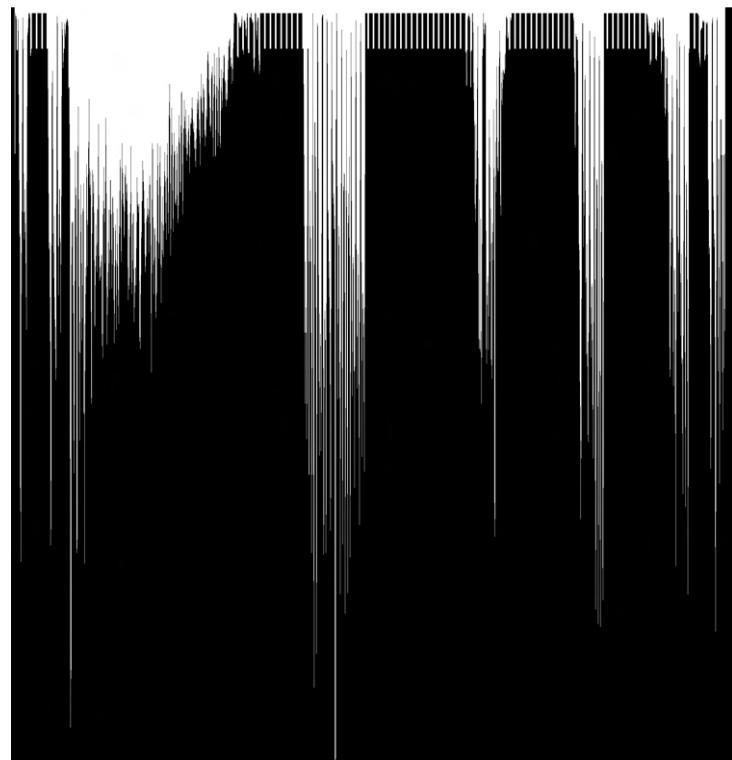
Rysunek 3.45: Surowa projekcja tabeli z rysunku 3.44 na oś poziomą



Rysunek 3.46: Sygnał z rysunku 3.45 przetransformowany i znormalizowany w zakresie wartości (0,1)

Lp.	Kod Nazwa towaru	Kod EAN	Ilość/j.m.	Cena netto	Wartość netto	VAT
1	447 Blok millimetrowy A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515 Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4 Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1 Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540 Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,8100	38,10	23%
6	45 Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,1700	190,20	23%
7	422 Blok szkolny w linie A4-100	5905824900190	20 szt.	2,2500	45,00	23%
8	41 Blok tech. A3-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42 Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14 Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412 Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93 Brulin 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99 Brulin 300k. A5 tw.opr.	5905824400546	2 szt.	7,4200	14,84	23%
14	139 Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3000	58,60	23%
15	170 Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856 Karton A3-20 złoty	5905824020386	2 opak.	8,5500	17,10	23%
17	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594 Mapa Europy -szkol.podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
19	595 Mapa Polski -szkol.podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599 Mapa Świata -szkol.podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501 Pap.fluo żółty samoprzylepny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476 Papier srebrny samoprzylepny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264 Papier (jk) A4-100 seleacyjny	5905824020249	1 opak.	4,7500	4,75	23%
24	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.47: Tabela kr1_0 poziomo obramowana ze zbioru treningowego



Rysunek 3.48: Surowa projekcja tabeli z rysunku 3.47 na oś poziomą



Rysunek 3.49: Sygnał z rysunku 3.48 przetransformowany i znormalizowany w zakresie wartości $(0,1)$

Segmentacja kolumnowa

Procedura wnioskowania dla rzutu na oś poziomą jest prosta i intuicyjna. Składają się na nią:

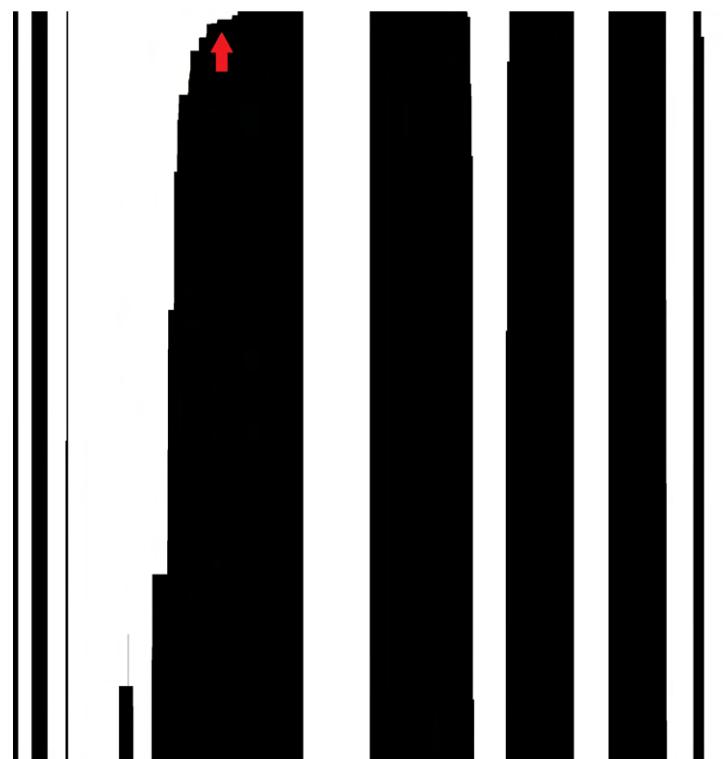
1. progowa detekcja dolin z wykorzystaniem parametru p_9
2. wyznaczenie miejsc segmentacji na podstawie rozkładu wartości w dolinach



Rysunek 3.50: Schemat wnioskowania segmentacji kolumnowej bez korzystania z obramowań

Dla każdej wykrytej doliny w jej wnętrzu ustalany jest punkt segmentacji. Początkowo, punkt ten był ustalany naiwnie po środku wykrytej przestrzeni.

Zauważono, że w przetransformowanych projekcjach kolumn często występują „ogony”, tzn. łagodniejsze zbocza zakończone bardzo małymi wartościami. Powodem ich występowania jest wyrównywanie tekstu do strony prawej lub lewej, przez co znaki po jednej ze stron pojawiają się znacznie rzadziej. Algorytm progowy czasami ignoruje to zjawisko, przez co zdarza się, że naiwnie ustalony punkt segmentacji przechodzi przez odstającą linię tekstu (rys. 3.52).



Rysunek 3.51: Przykład „ogona” kolumny, powstały na skutek wyrównywania tekstu

Lp.	Kod Nazwa towaru	Kod EAN	Ilość/j.m.	Cena netto	Wartość netto	VAT
1	447 Blok milimetrowy A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515 Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4 Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1 Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540 Blok rys. z kolorowymi kartkami Premium A3 30	5905824000463	10 szt.	3,8100	38,10	23%
6	45 Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,1700	190,20	23%
7	422 Blok szkolny w linie A4-100	5905824900190	20 szt.	2,2500	45,00	23%
8	41 Blok tech. A3-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42 Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14 Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412 Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93 Brulion 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99 Brulion 300k. A5 tw.opr.	5905824400546	2 szt.	7,4200	14,84	23%
14	139 Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3000	58,60	23%
15	170 Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856 Karton A3-20 złoty	5905824020386	2 opak.	8,5500	17,10	23%
17	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594 Mapa Europy -szkol.podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
19	595 Mapa Polski -szkol.podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599 Mapa Świata -szkol.podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501 Pap.fluo.żółty samoprzylepny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476 Papier srebrny samoprzylepny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264 Papier (jk) A4-100 seledynowy	5905824020249	1 opak.	4,7500	4,75	23%
24	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.52: Przykład błędnego wnioskowania spowodowanego nieuwzględnieniem ogona

Aby temu zapobiegać, postanowiono punkt segmentacji wewnętrz dolin ustalać na podstawie rozkładu wartości wewnętrz dolin, zgodnie z algorytmem:

1. punkty w dolinie dzielone są w punkcie środkowym na strony prawą i lewą
2. wartości punktów po obu stronach są porównywane jednostronnie testem t-Studenta z poziomem istotności $\alpha = 0.01$
3. w przypadku statystycznego wykazania większej średniej dla jednej ze stron, punkt segmentacji jest przesuwany o w 25% stronę przeciwną, w przeciwnym razie punkt pozostawiany jest na środku

Tak proste rozwiążanie okazało się wystarczające dla wielu przypadków w zbiorze treningowym. W niektórych tabelach algorytm dalej przecina pojedyncze wyrazy (rys. 3.53), wynika to też w dużej mierze z niedoskonałości przetwarzania wstępne (na rysunku nie widać ogonów ze względu na zbyt małą wartość sygnału po normalizacji). Widoczne są miejsca do wprowadzenia poprawek, choć błędy popełniane są głównie w nagłówkach tabel, a segmentacja ciała tabeli osiąga bardzo dobre, stabilne rezultaty.

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m ² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamentowa biel 20 szt./op. 220g/m ² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m ² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 125 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%
6	Folia laminacyjna PE arkusz A4 125 mic standard błysk antystatyczna Kod Indeksu: 320412	1	OP	29,50 PLN		29,50 PLN	29,50 PLN	23%
7	HEYKKA Zakreślacz klasyczny Linea jasnozielony, 10 szt./opk. Kod Indeksu: 611109	1	OP	21,00 PLN	35%	13,65 PLN	13,65 PLN	23%
8	HEYKKA Zakreślacz klasyczny Linea pastelowy żółty 10 szt./opk. Kod Indeksu: 611159	1	OP	21,00 PLN	35%	13,65 PLN	13,65 PLN	23%

Rysunek 3.53: Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sygnału z rysunku 3.46

L.p.	Kod	Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
1	447	Blok milimetrowy A4-20	5905824600878	20 szt.	1,1300	22,60	23%
2	515	Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,0100	100,50	23%
3	4	Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,0600	20,60	23%
4	1	Blok rys. kolorowy A4-16	5905824400966	10 szt.	1,0000	10,00	23%
5	540	Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,8100	38,10	23%
6	45	Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,1700	190,20	23%
7	422	Blok szkolny w linie A4-100	59058249000190	20 szt.	2,2500	45,00	23%
8	41	Blok tech. A3-10 250g	5905824801664	50 szt.	2,5700	128,50	23%
9	42	Blok tech. A4-10 250g	5905824801671	50 szt.	1,3200	66,00	23%
10	14	Blok techn. z czarnymi kartkami A4-10	5905824000043	10 szt.	1,4400	14,40	23%
11	412	Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,7300	111,90	23%
12	93	Brulion 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
13	99	Brulion 300k. A5 tw.opr.	5905824400546	2 szt.	7,4200	14,84	23%
14	139	Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3000	58,60	23%
15	170	Karton A1-20 czerwony	5905824200450	1 opak.	19,0000	19,00	23%
16	856	Karton A3-20 złoty	5905824020386	2 opak.	8,5500	17,10	23%
17	210	Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,5600	12,48	23%
18	594	Mapa Europy -szkol. podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
19	595	Mapa Polski -szkol. podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
20	599	Mapa Świata -szkol. podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
21	501	Pap.fluo.żółty samoprzylepny A4-20	5905824800575	1 opak.	8,7700	8,77	23%
22	476	Papier srebrny samoprzylepny A4-20	5905824120628	1 opak.	14,5200	14,52	23%
23	264	Papier (jk) A4-100 seledynowy	5905824020249	1 opak.	4,7500	4,75	23%
24	269	Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%

Rysunek 3.54: Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sygnału z rysunku 3.49

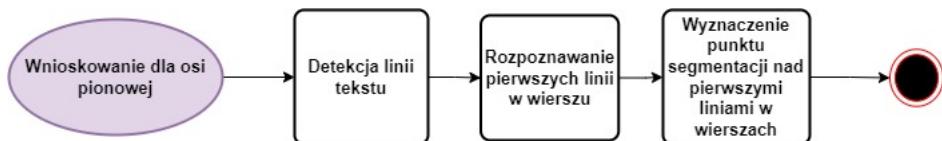
Segmentacja wierszowa

W osi pionowej procedura prezentuje się odmiennie. Jak wspominano, projekcja na oś pionową zazwyczaj złożona jest z słupków charakterystycznych dla linii tekstu. Podejście analogiczne co dla segmentacji kolumnowej, tzn. postawienie punktu segmentacji w każdej dolinie, skazane jest na niepowodzenie dla każdej tabeli, dla której komórki złożone są z więcej niż jednej linii tekstu.

Klasyfikacja dolin na szersze i węższe zgodnie z metodą Zuyeva również nie jest wystarczająca dla wielu przykładów z zaproponowanego zbioru. Często dla zaoszczędzenia przestrzeni, wiersze w tabelach są umieszczane jeden pod drugim, bez dodatkowej przestrzeni, a rozpoczęcie nowej pozycji jest sygnalizowane jedynie obecnością numeru w kolumnie indeksującej.

Przyglądając się pionowym projekcjom problematycznych tabel, zaobserwowano pewną zależność. Zwyczajowo, tekst w komórkach wyrównywany jest do góry, przez co dla pierwszej linii każdego wiersza wysokość słupka projekcji zdecydowanie przewyższa pozostałe słupki tego wiersza.

Spostrzeżenie to zainspirowało wykorzystanie pomysłu segmentowania wierszy tabeli poprzez rozpoznawanie linii tekstu rozpoczynających wiersz (podejścia podobnego do klasyfikacji BIESO [4]).



Rysunek 3.55: Schemat wnioskowania segmentacji wierszowej bez korzystania z obramowań

Do rozpoznawania wykorzystano jedynie jedną cechę - przybliżoną wysokość słupka tekstu na przetransformowanym obrazie projekcji. Wykonane operacje potęgowania pomagały odseparować progowo poszukiwane linie od pozostałych, warunkiem było dobranie odpowiednich wartości parametrów p_6 i p_7 . Obserwując uważnie wykres przetransformowanego sygnału nie stanowiło to większego problemu.

Otrzymane efekty były nadspodziewanie dobre. Algorytm działał bezbłędnie dla wszystkich faktur pionowo obramowanych w zbiorze treningowym z wykorzystaniem tylko jednego zestawu parametrów.

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

Lp	Symbol	EAN	Symbol obcy	Opis	JM	Ilosc	Cena PLN	Netto PLN	Vat %	Brutto PLN	
1	HA 3720 2030-M10	5905130007736		Blok malarski Młody Artysta, A4, 10 ark, 200g, Happy Color	szt	5	3,09	15,45	23%	3,55	19,00
2	HA AKPB1471-3	6933631545395		Długopis usuwalny "bużki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	2,63	31,56	23%	7,26	38,82
3	HA AKPB1474-3	6941025162721		Długopis usuwalny, seria Space/Story, 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	2,80	33,60	23%	7,73	41,33
4	HA AKPA6571-3	6941025110098		Długopis usuwalny "kaczki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	2,63	31,56	23%	7,26	38,82
5	HA AKPB4471-3	6941025125399		Długopis usuwalny "uszaki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	36	2,63	94,68	23%	21,78	116,46
6	HA AKPB7371-3	6941025125429		Długopis usuwalny "zyrafy", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	24	2,63	63,12	23%	14,52	77,64
7	FO 77894	4001868073010		Drut pluszowy MIX zielony, 10 szt, dt. 50cm, śr. 8mm, Folia	opk	2	3,29	6,58	23%	1,51	8,09
8	LA L300-95	4024526003501		Dziurkacz L300, 25 kartek, czarny, Laco	szt	1	15,52	15,52	23%	3,57	19,09
9	LA L300-25	4024526003525		Dziurkacz L300, 25 kartek, czarny, Laco	szt	1	15,52	15,52	23%	3,57	19,09
10	HA 7370 0075-75	5905130009333		Farba akrylowa 75ml, brązowy, Happy Color	szt	1	3,70	3,70	23%	0,85	4,55
11	HA 7370 0075-75	5905130007262		Farba akrylowa 75ml, ciemnobrązowy, Happy Color	szt	1	3,70	3,70	23%	0,85	4,55
12	S 317-9	4007817310748		Foliosis Lumocolor, M, wodoodporny, czarny, Staedtler	szt	10	4,28	42,80	23%	9,84	52,64
13	S 250 07-HB	4007817213889		Graffity 0,7 mm, na papier, HB, Staedtler	szt	12	2,19	26,28	23%	6,04	32,32
14	S 250 05-2B	4007817213820		Graffity 0,5 mm, na papier, 2B, Staedtler	szt	12	2,19	26,28	23%	6,04	32,32
15	S 250 05-HB	4007817213882		Graffity 0,5 mm, na papier, HB, Staedtler	szt	12	2,19	26,28	23%	6,04	32,32
16	LA H400-35	4024526000029		Zszywacz metalowy H400, 30 kartek, głębokość zszywania 64 mm, niebieski, Laco	szt	1	12,37	12,37	23%	2,85	15,22
17	S 144 NC12 SET6	4007817609019		Kredki szkolne Noris Club, 12 kolorów + ołówek + gumka, Staedtler	szt	2	6,92	13,84	23%	3,18	17,02
18	HA AKR67K35-3	6937168812296		Wkładki do długopisu usuwalnego, Standard A, 0,5mm, niebieski, 3 szt. w etui, Happy Color	szt	30	4,91	147,30	23%	33,88	181,18
19	S 120-HB	4007817104620		Ołówek Noris, HB, nr 2, Staedtler	szt	12	1,34	16,08	23%	3,70	19,78
20	S 350-2	4007817321485		Marker Lumocolor, wodoodporny, gruby, ścieśnia końcówka, czarny, Staedtler	szt	10	2,27	22,70	23%	5,22	27,92

Rysunek 3.56: Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sygnału z rysunku 3.40

lp.	mag	kod kreskowy	nazwa towaru/usługi	kod CN /PNWIU	j.n.	ilość	cena netto	wartość netto	kwota VAT	wartość brutto	masa (kg.)	
1	A01	5902277274458	BBI LINIJKA FLEX 15CM B&B KIDS PASTEL	SZT	20	0,83	16,60	23%	3,82	20,42	0,300	
2	A01	5902277060006	BLOK RYSUNKOWY A4 20 80G	SZT	40	0,84	33,60	23%	7,73	41,33	5,440	
3	A01	5902277070012	BLOK TECHNICZNY A3 10 170G	SZT	120	1,95	234,00	23%	53,82	287,82	31,200	
4	A01	5902277070005	BLOK TECHNICZNY A4 10 170G	SZT	120	0,97	116,40	23%	26,77	143,17	15,600	
5	A01	5902277170828	BRULION A4 96# M 70G	SZT	10	4,46	44,60	23%	10,26	54,86	6,100	
6	A01	5902277171726	BRULION A4 96# M 70G	SZT	5	4,46	22,30	23%	5,13	27,43	3,050	
7	A01	5902277170804	BRULION A5 96# M 70G	SZT	10	2,35	23,50	23%	5,41	28,91	2,920	
8	A01	5902277294241	BRULION A5 96# M 70G METALLIC SATIN GOLD	SZT	5	2,96	14,80	23%	3,40	18,20	1,460	
9	A01	5902277170606	BRULION A5 96# M 70G	SZT	10	2,24	22,40	23%	5,15	27,55	1,450	
10	A01	5902277179760	DZIENNIK KORESPONDENCYJNY A4 96 70G	SZT	5	7,36	36,80	23%	8,46	45,26	2,700	
11	A01	5902277295118	BBB OŁÓWEK B&B KIDS PASTEL	SZT	72	0,93	66,96	23%	15,40	82,36	0,432	
12	A01	5902277295262	BBB OŁÓWEK ZE ZWIERZAKIEM B&B	SZT	36	1,90	68,40	23%	15,73	84,13	0,504	
13	A01	5902277276551	INT ZAKĘ. INDEKSUJĄCE FUNKY 200 12X45MM	SZT	3	2,72	8,18	23%	1,88	10,04	0,033	
14	A01	5902277274625	KOLORONANKA Z NAKL.A4 16 DUŻO DO KOLOR.	58.11.13.0	SZT	10	2,06	20,60	5%	1,03	21,63	1,130
15	A01	5902277270418	KOŁOZ. PP A5 100# SPIR.PO RUR.BOKU Z G&B	SZT	5	5,81	29,05	23%	6,68	35,73	1,145	
16	A01	5902277171337	KOŁOZESZYT A5 160# M 70S Z PERF.	SZT	5	4,23	21,15	23%	4,86	26,01	1,750	
17	A01	5902277174437	KOSTKA PAP.8,5x8,5x3,5CM BIAŁA KLEJONA	SZT	30	1,25	37,50	23%	8,63	46,13	8,190	
18	A01	5902277170569	KOSTKA PAP.8,5x8,5x3,5CM KOLOR KLEJONA	SZT	60	1,63	97,80	23%	22,49	120,29	12,180	
19	A01	5902277178084	N-KOSTKA PAP.8,5x8,5x3,5CM KOLOR NIEKLEJ	SZT	10	1,63	16,30	23%	3,75	20,05	2,200	
20	A01	5902277244475	NSR KLEJ BROKATOWY NOSTER 6ml a'5	SZT	3	1,49	4,47	23%	1,03	5,50	0,195	
21	A01	5902277100009	PAPIER KOLOROWY A5 10 115G	SZT	10	0,67	6,70	23%	1,54	8,24	0,500	
22	A01	5902277275035	TDS DŁUGOPIS JET LINE BLACK a'10	SZT	10	0,99	9,90	23%	2,28	12,18	0,100	
23	A01	5902277275042	TDS DŁUGOPIS TODAYS JET LINE BLUE a'10	SZT	10	0,99	9,90	23%	2,28	12,18	0,100	
24	A01	5902277170118	WREKAD DO SEG A4 50# KOLM 70G	SZT	5	2,66	13,30	23%	3,06	16,36	1,175	
25	A01	5902277244543	YNI BROKAT SYPKI T6 A'6 PEARL	SZT	3	6,39	19,17	23%	4,41	23,58	0,315	
26	A01	5902277278203	INT DŁUGOPIS ZELOWE 6SZT. NEON	SZT	2	6,40	12,80	23%	2,94	15,74	0,024	
27	A01	5902277278210	INT DŁUGOPIS ZELOWE 6SZT. PASTEL	SZT	2	6,40	12,80	23%	2,94	15,74	0,024	
28	A01	5902277278197	INT DŁUGOPIS ZELOWE 6SZT.METALLIC	SZT	2	6,40	12,80	23%	2,94	15,74	0,220	
29	A01	5902277265531	DESEZYT A4 60# 70G PP	SZT	5	3,40	17,00	23%	3,91	20,91	1,500	
30	A01	5902277293862	N-ZESZYT A4 60# M 70G HS KRAFT	SZT	5	3,15	15,75	23%	3,62	19,37	1,460	
31	A01	5902277175014	ZESZYT A4 80# M 70G UV	SZT	5	3,14	15,70	23%	3,61	19,31	1,900	

KONIEC STRONY: 1

Rysunek 3.57: Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sygnału z rysunku 3.43

Wnioskowanie na podstawie tła

Niektóre tabele posiadały dodatkowy rodzaj wskazówek wizualnych informujących o miejscu segmentacji, mianowicie zmianę koloru tła pomiędzy sąsiadującymi komórkami. Przykładowo, nagłówek tabeli z rysunku 3.58 jest wyróżniony szarym tłem, jednak po binaryzacji 3.59 rozgraniczenie pomiędzy wierszami przestaje być widoczne.

Opracowano sposób ekstrakcji tych informacji z wykorzystaniem binaryzacji dwuprogowej i operacji morfologicznych. Tryb wyszukiwania tego typu wskazówek jest opcjonalny i można go aktywować przez parametr p_4 (choć nie zaobserwowano przypadków, gdy aktywowanie go wpłynęło negatywnie na wyniki segmentacji).

Wymaga on dodatkowego parametru p_5 informującego metodę o drugim progu binaryzacji, czyli odcieniu tła interesujących komórek (dobranemu eksperymentalnie). Metodologia jest następująca:

1. Obraz w skali szarości jest binaryzowany osobno z wykorzystaniem dwóch progów p_1 i p_5 (wynikami są obrazy bin_{p_1} i bin_{p_5}).
2. Obliczana jest różnica obrazów $bin_{dif} = bin_{p_5} - bin_{p_1}$.
3. Na obrazie bin_{dif} wykonywane są operacje erozji i dylatacji z dużym, kwadratowym elementem strukturalnym 20×20 ($R_1 = (bin_{dif} \ominus J_{20x20}) \oplus J_{20x20}$). Mają na celu wyeliminowanie szumów powstałych na skutek kompresji obrazu wejściowego i nieidealnie dobranej metody binaryzacji.
4. Otrzymanym wynikiem jest podziurawiony zarys komórek z poszarzonym tłem. W celu eliminacji dziur wykonywana jest operacja zamknięcia ($R_2 = (R_1(\oplus J_{20x20})^6)(\ominus J_{20x20})^6$)
5. Z otrzymanego pełnego zarysu poszukiwanych komórek pobierany jest obwód. W tym celu wykonywana jest dylatacja z jądrem 2×2 , od której wyniku jest odejmowany obraz przed dylatacją ($R_3 = R_2 \oplus J_{2x2} - R_2$)

W wyniku przetwarzania otrzymywane jest obramowanie poszukiwanych poszarzonych komórek (R_3). Jest ono addytywnie dołączane do pozostałych obramowań wykorzystywanych przy wnioskowaniu.

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 220g/m² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%

Rysunek 3.58: tabela FV210310097_0 z poszarzonym wierszem ze zbioru treningowego

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 220g/m² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%

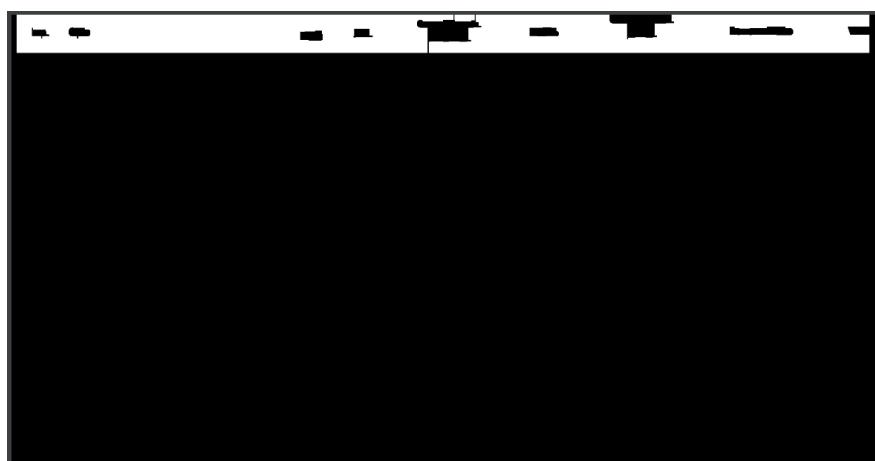
Rysunek 3.59: tabela z rysunku 3.58 po binaryzacji z progiem podstawowym

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gładki kremowy 20 szt./op. 250 g/m ² Kod Indeksu: 202802	23	OP	4,90 PLN	4,90 PLN	112,70 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamentowa biel 20 szt./op. 220g/m ² Kod Indeksu: 200604	12	OP	5,90 PLN	5,90 PLN	70,80 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium bławkitny 20 szt./op. 220g/m ² Kod Indeksu: 200708	12	OP	5,90 PLN	5,90 PLN	70,80 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN	24,00 PLN	120,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320400	3	OP	19,00 PLN	19,00 PLN	57,00 PLN	57,00 PLN	23%

Rysunek 3.60: tabela z rysunku 3.58 po binaryzacji z progiem dodatkowym

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Carton dekoracyjny kremowy 20 szt./op. 250 g/m ² Kod Indeksu: 202802	23	OP	4,90 PLN	4,90 PLN	112,70 PLN	112,70 PLN	23%
2	Carton dekoracyjny Island diamentowa biel 20 szt./op. 220g/m ² Kod Indeksu: 200604	12	OP	5,90 PLN	5,90 PLN	70,80 PLN	70,80 PLN	23%
3	Carton dekoracyjny Millennium bławkitny 20 szt./op. 220g/m ² Kod Indeksu: 200708	12	OP	5,90 PLN	5,90 PLN	70,80 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN	24,00 PLN	120,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320400	3	OP	19,00 PLN	19,00 PLN	57,00 PLN	57,00 PLN	23%

Rysunek 3.61: tabela z rysunku 3.58 po kroku 2. ($bin_{dif} = bin_{p5} - bin_{p1}$) przetwarzania



Rysunek 3.62: tabela z rysunku 3.58 po kroku 3. (usuwanie zaszumień) przetwarzania

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 220g/m² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%

Rysunek 3.63: tabela z rysunku 3.58 po kroku 4. (wypełnianie luk) przetwarzania

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 220g/m² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%

Rysunek 3.64: tabela z rysunku 3.58 po kroku 5. (ekstrakcja konturów) przetwarzania

L.p.	Opis	Ilość	J.M.	Cena jednostk. przed rab.	Rab%	Cena jednostk. po rab.	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/m² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	23%
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 220g/m² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
3	Karton ozdobny Millennium błękitny 20 szt./op. 220g/m² Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	23%
4	Folia laminacyjna PE arkusz A4 100 mic standard błysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	23%
5	Folia laminacyjna PE arkusz A4 80 mic standard błysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	23%

Rysunek 3.65: Wyniki ekstrakcji informacji z tła tabeli z rysunku 3.58

3.3.7 Łączenie wyników

Do punktów segmentacji otrzymanych na skutek przetwarzania bez obramowań dodawane są punkty z wykrytych obramowań. W wyniku tego może pojawić się powielenie punktów segmentacji w jednym miejscu i w efekcie puste wiersze / kolumny. Są one eliminowane w post - procesingu dwuetapowo:

1. Progowo sprawdzane jest, czy wiersz / kolumna zawiera tekst (analogicznie co przy modułach wnioskujących). W przypadku niewykrycia tekstu, pierwszy z punktów segmentacji jest usuwany.
2. Już po wydzieleniu komórek i zastosowaniu modułu OCR, kolumny / wiersze, w których nie wykryto tekstu są usuwane.

Otrzymane punkty jednoznacznie definiują proponowaną przez algorytm kratę tabeli. Na podstawie zidentyfikowanej kraty, z obrazu wejściowego wyekstrahowane zostają komórki, na każdej z nich osobno używany jest moduł OCR, rozpoznany tekst jest zapisywany w ramce danych, która na końcu zostaje przekonwertowana do CSV.

A	B	C	D	E	F	G	H	I
L.p.	Opis	Ilość	: J.M.	Zena jedne	Ra b/o/d	Cena jednc	Razem Netto	VAT
1	Karton ozdobny Gladki kremowy 20 szt./op. 250 g/ m ² Kod Indeksu: 202802	23	OP	4,90 PLN		4,90 PLN	112,70 PLN	230/0
2	Karton ozdobny Iceland diamantowa biel 20 szt./op. 2209] m ² Kod Indeksu: 200604	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	230/0
3	Karton ozdobny Millennium błękitny 20 szt. / op. ZZG/mz Kod Indeksu: 200708	12	OP	5,90 PLN		5,90 PLN	70,80 PLN	230/0
4	Folia laminacyjna PE arkusz A4 100 mic standard blysk antystatyczna Kod Indeksu: 320410	5	OP	24,00 PLN		24,00 PLN	120,00 PLN	230/0
5	Folia laminacyjna PE arkusz A4 80 mic standard blysk antystatyczna Kod Indeksu: 320480	3	OP	19,00 PLN		19,00 PLN	57,00 PLN	230/0
6	Folia laminacyjna PE arkusz A4 125 mic standard blysk antystatyczna Kod Indeksu: 320412	1	OP	29,50 PLN		29,50 PLN	29,50 PLN	230/0
7	HEYKKA Zakreślač klasyczny Linea jasnozielony, 10 szt.] opk. Kod Indeksu: 611109	1	OP	21,00 PLN	350/0	13,65 PLN	13,65 PLN	230/0
8	HEYKKA Zakreślač klasyczny Linea pastelowy żółty 10 szt.] opk. Kod Indeksu: 611159	1	OP	21,00 PLN	350/0	13,65 PLN	13,65 PLN	230/0

Rysunek 3.66: Wynikowy plik CSV dla tabeli z rys. 3.44

1	LP	Symbol	EAN	S	Opis	JM	IIOSC	Cena PLN	Netto PLN	%	Vat PLN	Brutto PLN
2	1	HA 3720 2030-M10	59051 30007736	Blok malarski Młody	szt	5	_ 3,09	15,45	0,23	3,55	19,00	
3	2	HA AKPB1471-3	6933631 545395	Dlugopis usuwalny "b"	SZt	12	_ 2,63	31,56	0,23	7,26	38,82	
4	3	HA AKPB1474-3	6941 0251 62721	Dlugopis usuwalny, se SZt		12	_ 2,80	33,60	0,23	7,73	41,33	
5	4	HA AKPA6571-3	6941 0251 1 0098	Dlugopis usuwalny "k"	SZt	12	_ 2,63	31,56	0,23	7,26	38,82	
6	5	HA AKPB4471-3	6941 0251 25399	Dlugopis usuwalny "u"	SZt	36	_ 2,63	94,68	0,23	21,78	116,46	
7	6	HA AKPB7371-3	6941 0251 25429	Dlugopis usuwalny "ż"	SZt	24	_ 2,63	63,12	0,23	14,52	77,64	
8	7	FO 77894	4001 86807301 O	Druty pluszowe Mix x opk		2	_ 3,29	6,58	0,23	1,51	8,09	
9	8	LA L300-95	4024526003501	Dziurkacz L300, 25 ka	SZt	1	_ 15,52	15,52	0,23	3,57	19,09	
10	9	LA L300-25	4024526003525	Dziurkacz L300, 25 ka szt		1	_ 15,52	15,52	0,23	3,57	19,09	
11	10	HA 7370 0075-7	59051 30006333	Farba akrylowa 75ml	SZt	1	_ 3,70	3,70	0,23	0,85	4,55	
12	11	HA 7370 0075-75	59051 30007262	Farba akrylowa 75ml	SZt	1	_ 3,70	3,70	0,23	0,85	4,55	
13	12	S 317-9	400781 731 0748	Folopis Lumocolor, N	SZt	10	_ 4,28	42,80	0,23	9,84	52,64	
14	13	S 250 O7-HB	400781 721 3889	Graffity 0,7 mm, na pa SZt		12	_ 2,19	26,28	0,23	6,04	32,32	
15	14	S 250 05-28	400781 721 3520	Graffity 0,5 mm, na pa SZt		12	_ 2,19	26,28	0,23	6,04	32,32	
16	15	S 250 05-HB	400781 721 3582	Graffity 0,5 mm, na pa SZt		12	_ 2,19	26,28	0,23	6,04	32,32	
17	16	LA H400-35	4024526000029	Zszywacz metalowy I	SZt	1	_ 12,37	12,37	0,23	2,85	15,22	
18	17	S 144 NC12 SET6	400781760901 9	Kredki szkolne Noris	I szt	2	_ 6,92	13,84	0,23	3,18	17,02	
19	18	HA AKR67K35-3	6937168812296	Wkłady do dlu90pisu	SZt	30	_ 4,91	147,30	0,23	33,88	181,18	
20	19	S 120-HB	400781 71 04620	Olówek Noris, HB, nr	SZt	12	_ 1,34	16,08	0,23	3,70	19,78	
21	20	S 350-2	4007817321485	Marker Lumocolor,	w SZt	10	_ 2,27	22,70	0,23	5,22	27,92	

Rysunek 3.67: Wynikowy plik CSV dla tabeli z rys. 3.38

A	B	C	D	E	F	G	H	I	J	K	L	M
In	n	kod kreskowy	nazwa towaru/usługi	kod CN /PKWII	j.m	ilość	cena netto	wartość netto	VAT	kwota vat	wartość brutto	masa (kg.)
2	I	AOI 5902272724458	BBI LINIUK FLEX 15CM BBB KIDS PASTEL	SZT	20	0.83	16.60	0,23	3.82	20.42	0.30	
3	2	AOI 5902277060006	BLOK RYSUNKOWY A4 20 BOG	SZT	40	0.84	33.60	0,23	7.79	41.39	5.440	
4	3	AOI 5902277070012	BLOK TECHNICZNY A4 10 NOG	SZT	120	1.95	234.00	0,23	53.82	287.82	91.200	
5	4	AOI 5902277070005	BLOK TECHNICZNY A4 10 17OG	SZT	120	0.97	116.40	0,23	26.77	143.17	15.600	
6	5	AOI 5902277170828	BRULION A4 96% M 70G	SZT	10	4.46	44.60	0,23	10.26	54.86	6.100	
7	6	AOI 5902277171726	BRULION A4 96% M 70G	SZT	5	4.46	22.30	0,23	5.13	27.43	3.050	
8	7	AOI 5902277170804	BRULION A4 96% M 70G	SZT	10	2.35	23.50	0,23	5.41	28.91	2.920	
9	8	AOI 5902277259421	BRULION A5 96% M 70G METALLIC SATIN GOLD	SZT	5	2.96	14.80	0,23	3.40	18.20	1.460	
10	9	AOI 5902277170606	BRULION A6 96% M 70G	SZT	10	2.24	22.40	0,23	5.15	27.55	1.450	
11	10	AOI 5902277179760	DZIENNIK KORESPONDENCYNY A4 % 70G	SZT	5	7.36	36.80	0,23	8.46	45.26	2.700	
12	II	AOI 5902277295118	BB OLÓWEK & B B KIDS PASTEL	SZT	72	0.93	66.96	0,23	15.40	82.36	0.432	
13	12	AOI 5902277295262	BB OLÓWEK ZE ZWIERZAKIEM B&B	SZT	36	1.90	68.40	0,23	15.73	84.13	0.504	
14	13	AOI 5902277276551	INT ZAKL.IDEKSUJĄCE FUNKY 200 12X45MM	SZT	3	2.72	8.16	0,23	1.88	10.04	0.033	
15	14	AOI 5902277274625	KOLOROWANKA 2 NAKL.A4 16 DUZO DO KOLOR. 58.11.13.0	SZT	10	2.06	20.60	0,05	1.03	21.63	1.130	
16	15	AOI 5902277207418	KOLOZ.PP AM 100% SPIR.PO KR.BOKU Z G&B	SZT	5	5.81	29.05	0,23	6.68	35.73	1.145	
17	16	AOI 5902277171337	KOLOZESZT A5 160# M 7% Z PERF.	SZT	5	4.23	21.15	0,23	4.86	26.01	1.750	
18	17	AOI 5902277174437	KOSTKA PAP 8,5X8,5X3,5CM BIAŁA KLEJONA	SZT	30	1.25	37.50	0,23	8.63	46.13	8.190	
19	18	AOI 5902277170569	KOSTKA PAP 8,5X8,5X3,5CM KOLOR KLEJONA	SZT	60	1.63	97.80	0,23	22.49	120.29	12.180	
20	19	AOI 5902277170804	N-KOSTKA PAP 8,5X8,5X3,5CM KOLOR NIELEJ	SZT	10	1.63	16.30	0,23	3.75	20.05	2.200	
21	20	AOI 5902277144475	NSR KLEJ BROKATOWY NOSTER 6ml a' 5	SZT	3	1.49	4.47	0,23	1.03	5.50	0.195	
22	21	AOI 5902277100009	PAPIER KOLOROWY A5 10 115G	SZT	10	0.67	6.70	0,23	1.54	8.24	0.500	
23	22	AOI 5902277275035	TOD DLUGOPIS TODAYS JET LINE BLACK a' 10	SZT	10	0.99	9.90	0,23	2.28	12.18	0.100	
24	23	AOI 5902277275042	TOD DLUGOPIS TODAYS JET LINE BLUE a' 10	SZT	10	0.99	9.90	0,23	2.28	12.18	0.100	
25	24	AOI 5902277170118	WKŁAD DO SFG A4 50% KOL.M 70G	SZT	5	2.66	13.30	0,23	3.06	16.36	1.175	
26	25	AOI 5902277244543	YNU BROKAT SPYKI 7G A'6 PEARL	SZT	3	6.39	19.17	0,23	4.41	23.58	0.315	
27	26	AOI 5902277278203	YNT DLUGOPIS ZELOWE 65ZT. NEON	SZT	2	6.40	12.80	0,23	2.94	15.74	0.024	
28	27	AOI 5902277278210	YNT DLUGOPIS ZELOWE 65ZT. PASTEL	SZT	2	6.40	12.80	0,23	2.94	15.74	0.024	
29	28	AOI 5902277278197	YNT DLUGOPIS ZELOWE 65ZT. METALLIC	SZT	2	6.40	12.80	0,23	2.94	15.74	0.220	
30	29	AOI 5902277265531	ZESZYT A4 60% M 70G PP	SZT	5	3.40	17.00	0,23	3.91	20.91	1.500	
31	30	AOI 5902277259862	N-ZESZYT A4 60% M 70G HS KRAFT	SZT	5	3.15	15.75	0,23	3.62	19.37	1.460	
32	31	AOI 5902277175014	ZESZYT A4 80% M 70G UV	SZT	5	3.14	15.70	0,23	3.61	19.31	1.900	

Rysunek 3.68: Wynikowy plik CSV dla tabeli z rys. 3.41

A	B	C	D	E	F	G	H
'Lp.	Ko	1 Nazwa towaru	Kod EAN	Ilość/Jm.	Ce	War	VAT
1	1	Blok milimetrowy A4-20	5905824600878	20 szt.	1,13	22,6	0,23
3	2	Blok Premium techniczny kolorowy A4-15	5905824020324	50 szt.	2,01	100,5	0,23
4	3	Blok rys. kolorowy A3-16	5905824400973	10 szt.	2,06	20,6	0,23
5	4	Blok rys. kolorowy A4-16	5905824400966	10 szt.	1	10	0,23
6	5	Blok rys. z kolorowymi kartkami Premium A3-30	5905824000463	10 szt.	3,81	38,1	0,23
7	6	Blok Superior techniczny kolorowy A4-25	5905824100859	60 szt.	3,17	190,2	0,23
8	7	Blok szkolny w linie A4-100	5905824900190	20 szt.	2,25	45	0,23
9	8	Blok tech. A3-10 250g	5905824801664	50 szt.	2,57	128,5	0,23
10	9	Blok tech. A4-10 250g	5905824801671	50 szt.	1,32	66	0,23
11	10	Blok techn., z czarnymi kartkami A4-10	5905824000043	10 szt.	1,44	14,4	0,23
12	11	Blok techniczny kolorowy ELITE A4-30	5905824501281	30 szt.	3,73	111,9	0,23
13	12	Brunil 300k. A4 tw.opr.	5905824000432	2 szt.	12,82	25,64	0,23
14	13	Brunil 300k. A5 tw.opr.	5905824400546	2 szt.	7,42	14,84	0,23
15	14	Kalendarz VITO A4 chamois	5905824020669	2 szt.	29,3	58,6	0,23
16	15	Karton A1-20 czerwony	5905824200450	1 opak.	19	19	0,23
17	16	Karton A3 -20 złoty	5905824020386	2 opak.	8,55	17,1	0,23
18	17	Kostka kolor 8,5x8,5 mix klejona	5905824000586	8 szt.	1,56	12,48	0,23
19	18	Mapa Europy -szkol.podkt.na biurko-intro	5905824400799	1 szt.	5,15	5,15	0,05
20	19	Mapa Polski -szkol.podkt.na biurko-intro	5905824200047	1 szt.	4,13	4,13	0,05
21	20	Mapa Świata -szkol.podkt.na biurko-intro	5905824400782	1 szt.	5,15	5,15	0,05
22	21	Pap.fluo.żółty samoprzyklejny A4-20	5905824800575	1 opak.	8,77	8,77	0,23
23	22	Papier srebny samoprzyklejny A4-20	5905824120628	1 opak.	14,52	14,52	0,23
24	23	Papier (jk) A4-100 seledynowy	5905824020249	1 opak.	4,75	4,75	0,23
25	24	Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,09	25,45	0,23

Rysunek 3.69: Wynikowy plik CSV dla tabeli z rys. 3.47

3.4 Porównywane metody ekstrakcji tabel

Dla rzetelnego określenia jaka faktycznie jest jakość końcowego systemu nie wystarczyło suchego wyliczenie metryk ewaluacyjnych osiągniętych na zbiorze testowym. Niezbędne było zestawienie wyników z innymi, podobnymi systemami, ocenianymi według tych samych kryteriów na tych samych danych. W tym celu wybrano dwa publicznie dostępne algorytmy, które pozwalały na przyjęcie danych z zaproponowanego zbioru i zwrócenie porównywalnych wyników.

Camelot

Pierwszy z algorytmów został zaimplementowany w ramach biblioteki open - source o nazwie Camelot ([2]). Rozwiązanie to było dostosowane jedynie do przetwarzania plików PDF z zawartymi metadanymi. Uznano, jednak że metoda ta może zostać wzięta pod uwagę w zestawieniu, ze względu na dostępność odpowiednich plików w zbiorze i możliwość wykorzystania efektów badań przy wdrożeniu (wszystkie faktury w zbiorze są udostępniane przedsiębiorstwu w formie generowanych cyfrowo PDF - ów, dlatego biblioteka mogła zostać praktycznie wykorzystana). Przewaga, jaką dają metadane przy ekstrakcji słów została częściowo zniwelowana poprzez dobór odpowiednich metod ewaluacji.

By rozwiązać problem detekcji tabeli (który w ramach autorskiej metody był pominięty poprzez bezpośredni odczyt współrzędnych z ground - truth) skorzystano z modułów detekcji udostępnionych w ramach biblioteki. By wyrównać szanse, w wypadku niedokładnej detekcji koordynaty tabeli były manualnie korygowane.

Camelot [2] jest rozwiązaniem w dużym stopniu sparametryzowanym. Można używać go w dwóch trybach:

- tryb stream:
 - metoda bottom-up
 - liczba kolumn jest zgadywana przy pomocy mody liczby słów w linii tekstu
 - słowa są przypisywane do kolumn na podstawie marginesów
 - wiersze są segmentowane naiwnie poprzez parametr grupujący *row_tol*
- tryb lattice:
 - metoda top - down wykorzystująca konwersję dokumentu do postaci obrazu
 - działa dla metod w pełni obramowanych
 - bliska standardowemu podejściu wykorzystującemu morfologię [1], [27]
 - wykorzystuje parametr *line_scale* do dopasowania długości elementu strukturyzującego cego

Założenia dotyczące dostrajania tej metody były analogiczne co dla algorytmu autorskiego. Dla każdego dostawcy dostrojono parametry algorytmu (tryb, *line_scale*, *row_tol*) do stylu faktury ze zbioru treningowego. Przyjęto założenie, że jeżeli manipulacja parametrami przestawała przynosić pozytywne rezultaty albo poprawa wyników w jednym obszarze tabeli skutkowała pogorszeniem w innym, strojenie dobierało końca, a parametry były zapisywane.

Po dostrojeniu na zbiorze treningowym dla wszystkich tabel testowych wygenerowane zostały pliki CSV, które oczekiwali na ewaluację.

Nanonets

Publicznie dostępne informacje o Nanonets [24] są ograniczone, co jest zrozumiałe, ze względu na jego charakter komercyjny.

Wiadomym jest jedynie, że ogranicza rozpoznawanie struktury tabeli do ustalenia kraty tabeli, a producenci chwalą się, że ich algorytmy wykorzystują zaawansowane modele głębokich sieci neuronowych. Model zwracanych wyników pozwala snuć pewne domysły odnośnie szczegółów rozwiązania, nie są to jednak rzetelne informacje.

Produkt cieszy się dużą popularnością, zbiera pozytywne recenzje i korzystają z niego duże firmy (m.in. AXA, Deloitte), dlatego został wybrany do zestawienia jako swego rodzaju wyznacznik jakości.

Producenci pozwalają na douczanie modelu z wykorzystaniem własnych danych przy kupieniu odpowiedniej licencji. Jak pokazały testy na zbiorze treningowym, nie było to jednak konieczne - nie zaobserwowano obecności innych błędów niż tych spowodowanych złożonością tabel. Te problemy nie mogły zostać rozwiązane bez zmiany fundamentalnych założeń modelu. Podobnie co w przypadku Camelot, rezultaty działania na obrazach tabel testowych zostały zapisane do formatu CSV.

3.5 Wybrane metody ewaluacji

Wybór wymiernej metody ewaluacji dopasowanej do wszystkich zestawianych systemów okazał się sporym wyzwaniem. W pierwszej kolejności rozważono wykorzystanie dokładnej implementacji jednej z w pełni automatycznych metod oceniania wykorzystywanych w literaturze.

Pod uwagę brano jedynie najpopularniejsze, sprawdzone algorytmy wykorzystane przy ocenie prac konkursowych - ICDAR2013 i ICDAR2019 [8], [16]. Uznano, że takie podejście będzie bardziej wartościowe od ewaluacji wykorzystywanych sporadycznie przez pojedyncze prace.

Metoda ICDAR2019 dopasowana do metod opartych na wizji komputerowej była ciekawą opcją pod względem ewaluacji metody autorskiej, gdyż segmentację wierszowo - kolumnową dość łatwo można przenieść do postaci bounding - boxów i ich relacji sąsiedztwa. Niestety, zarówno Camelot, jak i Nanonets nie zwracają informacji o segmentacji obrazu - pozwalają jedynie na wygenerowanie końcowego pliku, takiego jak CSV. Było to powodem zrezygnowania z tego podejścia.

Metoda ICDAR2013 była bardziej realna w kontekście oceny zestawianych algorytmów TSR. Identyfikacja komórek w ground-truth miałaby przebiegać poprzez dokładne porównywanie zawartości tekstu co możliwe byłoby również dla plików CSV. Stanowiło to również pewną przeszkodę - porównywanie tekstu nie jest do końca sprawiedliwe w stosunku do metod wykorzystujących OCR (autorskiej i Nanonets). Oprócz tego, w metodzie ignorowane są komórki puste, których prawidłowa identyfikacja jest istotnym wynikiem działania systemu.

Wybrano rozwiązanie będące kompromisem pomiędzy automatyzacją, a rzetelnością i sprawiedliwością oceny. Rozwiązanie te miało dać użytkownikowi możliwość samodzielnego odróżniania i ignorowania błędów spowodowanych działaniem OCR od istotnych z punktu widzenia algorytmu błędów segmentacji. Manualne podejście miało pozwolić również na traktowanie pustych komórek na takich samych zasadach co komórki tekstowe. Przyjmując te założenia, wszystkie pozostałe reguły miały być identyczne do tych wykorzystanych w ICDAR2013.

W celu realizacji założeń ewaluacji sporządzono narzędzie do manualnego tagowania połączeń pomiędzy komórkami w pliku CSV. Użytkownik przez odznaczenie checkboxa symbolizującego połaczanie pomiędzy sąsiadami mógł sklasyfikować je jako błędne, następnie program sam wyliczał metryki (precyję, czułość, f1 - score) na podstawie pobranych danych z połączeń i wcześniej sporzązonego ground - truth, zawierającego informacje o liczbie kolumn i wierszy w każdej tabeli.

Ocena zgodna z ICDAR2013 była realizowana w prosty sposób - wszystkie połączenia wychodzące od komórek, w których zaobserwowano wyniki błędnej segmentacji (dodatkowe znaki, brak istotnych znaków, przekłamania spowodowane ucięciem znaku) były klasyfikowane jako błędne. Przygotowanie wyników wymagało jedynie minimalnej interpretacji osoby weryfikującej przy klasyfikacji błędów na spowodowane przez OCR lub segmentację (zazwyczaj klasyfikacja ta była dość jednoznaczna).

Niestety, metoda ewaluacji czerpiąca z ICDAR2013 ma swoje słabe strony. W przypadku wystąpienia regularnej, nadmiernej lub niewystarczającej segmentacji w jednej z osi (przeważnie pionowej), często wyniki końcowe były bardzo słabe (na poziomie 0.2 f1 - score) pomimo perfekcyjnej segmentacji w drugiej osi (rys. 3.74).

Postanowiono wprowadzić modyfikację do metody ICDAR2013, tak by była mniej restrykcyjna i nie karała połączeń za błędną segmentację po drugiej stronie komórki lub w innej osi (rys. 3.75).

Zasady opracowanej, zmodyfikowanej ewaluacji prezentowały się następująco:

- jeśli w komórce znajdował się tekst, który powinien znajdować się w sąsiedniej komórce lub występowały jasne znaki, że segmentacja pomiędzy tymi komórkami była błędna (np. ucięcie znaku), połączenie pomiędzy komórkami było uznawane za błędne
- wszelkie połączenia pomiędzy dodatkowymi kolumnami / wierszami powstały poprzez nadmierną segmentację były uznawane za błędne (miało to na celu zapobieganie wyliczaniu czułości przekraczającej 1)
- Nieprawidłowe zespolenia kolumn / wierszy były ignorowane, z założeniem, że odpowiedni błąd został naliczony w metryce czułości

W efekcie opracowana metoda była mniej restrykcyjną odmianą metody wykorzystanej w ICDAR2013 - wszystkie połączenia sklasyfikowane jako błędne w modyfikacji, były błędne również w oryginale, ale nie odwrotnie.

Przykładowo, dla tabeli z rys. 3.72, dla której segmentacja kolumnowa metody Camelot była perfekcyjna, metoda oryginalna wskazała f1 - score o wartości 0.2602 (precyzja 0.1940, czułość 0,3950). Modyfikacja ICDAR2013 wskazała natomiast f1 - score na poziomie 0.6543 (precyzja 0.4879, czułość 0.9933). Metoda Camelot dla przedstawionego przykładu zastosowała częstą, nadmierną segmentację co powinno skutkować wskazaniu małej precyzji i niemal idealnej czułości. Pokazuje to, że zaproponowana modyfikacja miała sens i mogła służyć jako uzupełnienie oryginalnej metody oceniania, choć w zaproponowanej półautomatycznej formie wymagała dużego zaangażowania osoby weryfikującej przy interpretowaniu wyników.

Lp.	Kod	Nazwa towaru	Kod EAN
1	447	Blok milimetrowy A4-20	5905824600878

Rysunek 3.70: Fragment tabeli kr_1 ze zbioru treningowego

'Lp.	<input type="checkbox"/>	Ko	<input type="checkbox"/>	1 Nazwa towaru	<input type="checkbox"/>	Kod EAN
	<input checked="" type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>	<input checked="" type="checkbox"/>
1	<input checked="" type="checkbox"/>	447	<input checked="" type="checkbox"/>	Blok milimetrowy A4-20	<input checked="" type="checkbox"/>	5905 824600878

Rysunek 3.71: Ewaluacja metody autorskiej na fragmencie tabeli z rys. 3.70 przy pomocy metody ICDAR2013

'Lp.	<input checked="" type="checkbox"/>	Ko	<input type="checkbox"/>	1 Nazwa towaru	<input checked="" type="checkbox"/>	Kod EAN
	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
1	<input checked="" type="checkbox"/>	447	<input checked="" type="checkbox"/>	Blok milimetrowy A4-20	<input checked="" type="checkbox"/>	5905 824600878
	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Rysunek 3.72: Ewaluacja metody autorskiej na fragmencie tabeli z rys. 3.70 przy pomocy metody zmodyfikowanej

Lp	Symbol	EAN	Symbol obcy	Opis	JM	Ilość	Cena PLN	Netto PLN
1	HA 3820 2030-MS	5905130037153		Blok SILVER, A4, 150-230g, 10 ark, Happy Color	szt	1	8,24	8,24
2	HA AKPB1471-3	6933631545395		Długopis usuwalny "buźki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	10	2,80	28,00
3	HA AKPB1474-3	6941025162721		Długopis usuwalny, seria Space, 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	3,04	36,48
4	HA AKPB7371-3	6941025125429		Długopis usuwalny "żyrafy", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	10	2,80	28,00

Rysunek 3.73: Fragment tabeli 2020_FA_MG01_14153_0 ze zbioru treningowego

Lp	Symbol	EAN	Symbol obcy	Opis	JM	Ilość	Cena PLN	Netto PLN
1	HA 3820 2030-MS	5905130037153		Blok SILVER, A4, 150-230g, 10 ark, Happy Color	szt	1	8,24	8,24
2	HA AKPB1471-3	6933631545395		Długopis usuwalny "buźki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	10	2,80	28,00
3	HA AKPB1474-3	6941025162721		Długopis usuwalny, seria Space, 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	3,04	36,48
4	HA AKPB7371-3	6941025125429		Długopis usuwalny "żyrafy", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	10	2,80	28,00

Rysunek 3.74: Ewaluacja Camelot autorskiej na fragmencie tabeli z rys. 3.73 przy pomocy metody ICDAR2013

Lp	Symbol	EAN	Symbol obcy	Opis	JM	Ilość	Cena PLN	Netto PLN
1	HA 3820 2030-MS	5905130037153		Blok SILVER, A4, 150-230g, 10 ark, Happy Color	szt	1	8,24	8,24
2	HA AKPB1471-3	6933631545395		Długopis usuwalny "buźki", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	10	2,80	28,00
3	HA AKPB1474-3	6941025162721		Długopis usuwalny, seria Space, 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	12	3,04	36,48
4	HA AKPB7371-3	6941025125429		Długopis usuwalny "żyrafy", 0,5mm, niebieski, kolorowa obudowa, Happy Color	szt	10	2,80	28,00

Rysunek 3.75: Ewaluacja metody Camelot na fragmencie tabeli z rys. 3.73 przy pomocy metody zmodyfikowanej

3.6 Zestawienie i analiza wyników

Otrzymane rezultaty na zbiorze testowym były porównywane dwuetapowo. W pierwszej kolejności zestawiono wyniki zbiorowe na wszystkich 63 testowych tabelach poprzez wyliczenie średnich i odchyleń standardowych wartości precyzji, czułości i f1 - score dla 3 zestawianych metod (Camelot, Nanonets i metody autorskiej, dalej nazywanej ProjectionP). Następnie pogrupowano tabele ze zbioru testowego ze względu na styl tabeli (dostawcę, od którego pochodził dokument źródłowy) i powtórzono zestawienie wewnątrz grup. Taka szczegółowość dokonanych analiz miała posłużyć zarówno identyfikacji mocnych i słabych stron każdej z metod (w szczególności metody autorskiej), jak i ocenie które metody i w jakim stopniu nadają się do przetwarzania konkretnych klas faktur w warunkach praktycznych.

	Camelot		ProjectionP*		Nanonets	
	ICDAR2013 eval.	Modified* eval.	ICDAR2013 eval.	Modified* eval.	ICDAR2013 eval.	Modified* eval.
avg. precision	0,7346	0,8525	0,929	0,9787	0,9477	0,9849
std. precision	0,2843	0,202	0,1491	0,0576	0,102	0,0306
avg. recall	0,786	0,9567	0,9063	0,9478	0,94	0,974
std. recall	0,2144	0,0608	0,1793	0,1003	0,125	0,0645
avg. f1-score	0,7539	0,8906	0,9163	0,9611	0,9432	0,9785
std. f1-score	0,259	0,1468	0,1656	0,0766	0,1141	0,0459

Tablica 3.6: Wyniki osiągnięte na pełnym zbiorze testowym przez zestawiane metody

Osiągnięte przez autorską metodę oszacowania prezentują się obiecująco na tle pozostałych wyników. Pod względem metodologii oceniania ICDAR2013, osiągnięta przez ProjectionP uśredniona miara f1 na poziomie 0.91 jest zbliżona do często cytowanych rozwiązań z literatury [29], [9], [25]. Należy pamiętać jednak, że wyniki te nie są bezpośrednio porównywalne - ewaluację przeprowadzono na różnych zbiorach, metoda autorska jest metodą parametryzowaną, a algorytmy DeepDeSRT [29] i TableNet [25] operujące na obrazach najpewniej były weryfikowane z penalizacją błędów spowodowanych działaniem OCR. Nie zmienia to jednak faktu, że ProjectionP w ramach zestawienia przygotowanego przez autora poradziła sobie bardzo dobrze, przewyższając bibliotekę Camelot o 0.16 f1-score.

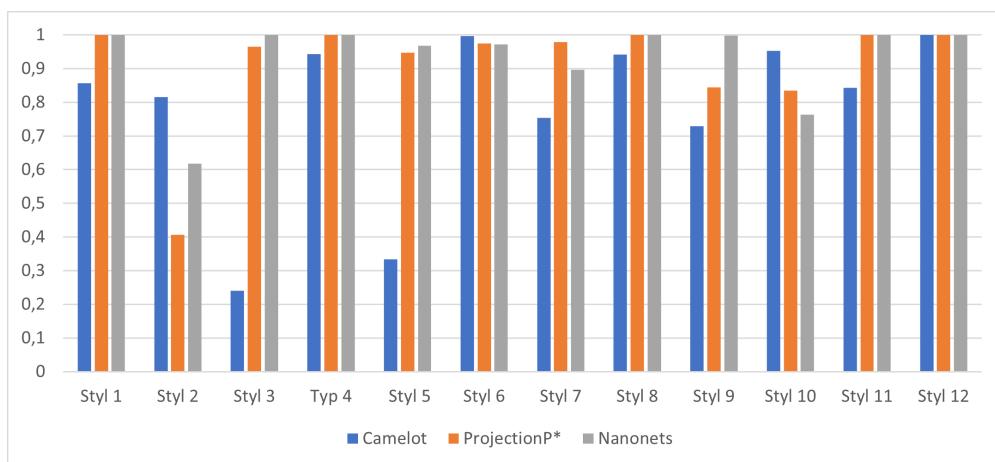
Należy podkreślić fakt, że Camelot została dostrojona tak, by czułość przewyższała precyzję i w wielu przypadkach można było zaobserwować dokonaną przez nią nadmierną segmentację. Manipulując parametrami możliwe było poprawienie precyzji kosztem czułości i niewykluczone, że miałyby to wpływ na końcową wartość miary f1. Teoretycznie mogłyby to poprawić wyniki o kilka setnych, ale z pewnością nie pozwoliłyby dorównać do metody autorskiej, która działała wyraźnie lepiej i stabilniej.

Z drugiej strony, ProjectionP nie dorównywało Nanonets, które działało niemal doskonale przy wyznaczaniu kraty tabeli we wszystkich przypadkach, a błędy były generowane głównie

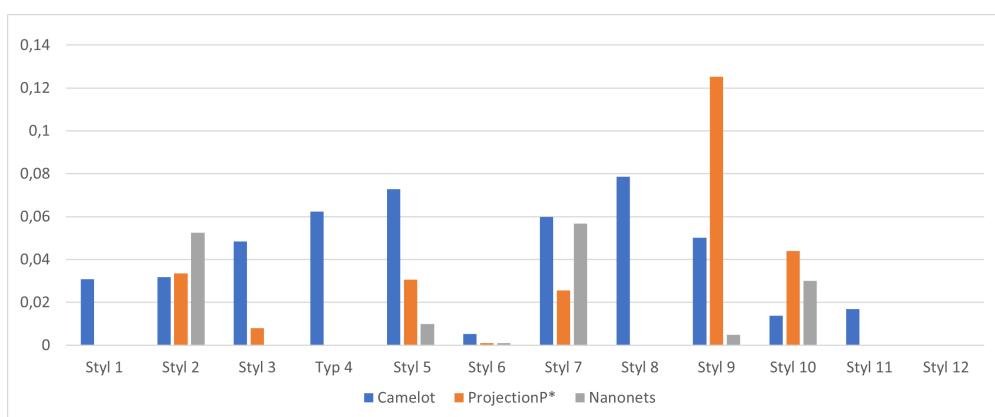
przez zespolone komórki, z którymi nie miało możliwości sobie poradzić ze względu na fundamentalne założenia.

Wprowadzenie modyfikacji do metody oceniania przybliżyło do siebie wyniki osiągnięte przez porównywane algorytmy TSR. Szczególnym beneficjentem była metoda Camelot, której udało się przewyższyć metodę autorską pod względem czułości (precyzja i ogólna dokładność pozostały zdecydowanie słabsze). Wydaje się, że tak wysoka ocena czułości osiągnięta przez Camelot (0.95) jest uzasadniona, gdyż większość rzeczywistych miejsc podziału tekstu zostawała w jej wypadku wykryta, a głównym trapiącym ją problemem była nadmierna segmentacja.

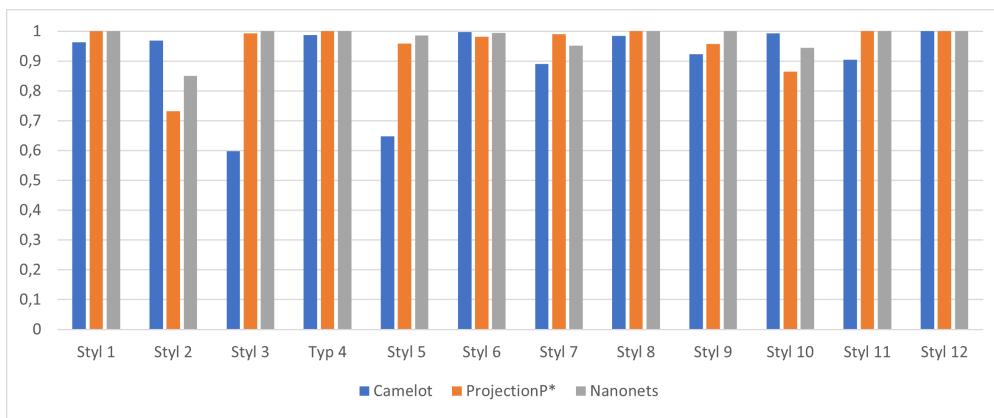
Martwiące pod kątem praktycznego zastosowania wydawały się wysokie wartości odchylenia standardowego f1 - score każdej z trzech metod (przekraczające 0.1). System wykorzystujący wyniki zestawianych algorytmów powinien być niezawodny, przez co powinny one zwracać dobre, stabilne wyniki. Przyczyny występowania tych odchyлеń zostały pokazane w dalszych analizach.



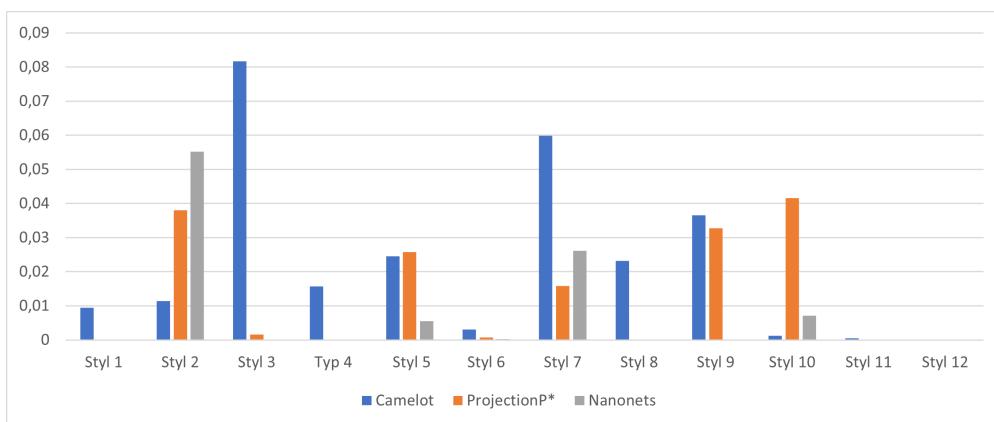
Rysunek 3.76: Zestawienie uśrednionego f1 - score obliczonego metodą ewaluacji ICDAR2013 z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)



Rysunek 3.77: Zestawienie odchylenia standardowego f1 - score obliczonego metodą ewaluacji ICDAR2013 z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)



Rysunek 3.78: Zestawienie uśrednionego f1 - score obliczonego zmodyfikowaną metodą ewaluacji z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)



Rysunek 3.79: Zestawienie odchylenia standardowego f1 - score obliczonego zmodyfikowaną metodą ewaluacji z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)

Ponownie wyliczono miary f1 - tym razem osobno agregując wszystkie tabele z dokumentów źródłowych należące do różnych dostawców dokumentu źródłowego (tym samym minimalizując różnorodność stylów w grupie). Jak pokazują rysunki 3.76 i 3.78, dla każdego z algorytmów można przyporządkować style tabel, dla których sprawuje się najlepiej oraz takie, z którymi sobie nie radzi. Bez zaskoczenia, najbardziej zawodny okazał się Camelot - w żadnym stopniu nie radził sobie ze stylami trzecim i piątym (do których należą tabele z rysunków 3.5 i 3.2). Powodem tak słabych wyników była naiwna agregacja linii tekstu, przez którą segmentacja wierszowa nie radziła sobie z komórkami kilkuliniowymi.

Zaskoczeniem były natomiast wyniki Camelot osiągnięte dla stylu drugiego (rys. 3.4). Były one zdecydowanie wyższe od wyników zdobytych przez ProjectionP i Nanonets. Metoda bazująca na pdf, korzystając z metadanych zawartych w pliku, była w stanie poprawnie pogrupować tekst w odpowiednich komórkach, nawet w miejscach, gdzie wychodził on poza fizyczny zakres kolumny

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

(rys. 3.81). Z powodów fundamentalnych ograniczeń, obie metody bazujące na obrazach nie były w stanie sprostać temu problemowi.

Lp.	Symbol Nazwa produktu	Ilość	J.m.	Rabat %	Cena netto	Kwota Netto	Podatek VAT		Kwota Brutto
							%	Kwota	
1	127321 FC GUMKA ARTYSTYCZNA CHLEBOWA MIX KOLETUPLASTIKOWE CZERWONA/NIEBIESKA/ZŁOTA FABER-CASTELL [127321 FC]	9556089009232	18 szt	20,00	2,15	38,70 P23	8,90		47,60
<hr/>									
2	300/001/C ED MARKER EDDING PERMANENTNY OKRĄGLA KONCÓWKA 1,5-3MM CZARNY [300/001/C ED]	4004764390564	10 szt	20,00	2,07	20,70 P23	4,76		25,46
3	54410-21063 DA NOŻYCZKI DAHLE COMFORT-GRIP, DO PAPIERU 24,7 cm [54410-21063 DA]	4007885232898	10 szt		7,91	79,10 P23	18,19		97,29
4	54405-21092 DA NOŻYCZKI DAHLE COMFORT-GRIP, DOMOWE 14 cm [54405-21092 DA]	4007885232867	10 szt		3,23	32,30 P23	7,43		39,73
5	117201 FC OŁÓWEK GRIP 2001 B Z GUMKĄ FABER-CASTELLCASTELL [117201 FC]	4005401172017	12 szt	20,00	2,15	25,80 P23	5,93		31,73
6	114000 FC OŁÓWKI GOLD FABER 1222 6 SZT. +GUMKA+TEMPERÓWKA OPAKOWANIE KARTONOWE FABER-CASTELL [114000 FC]	4005401140009	1 kpl	20,00	17,59	17,59 P23	4,05		21,64
7	59408-00000-00 TS PŁATKI SAMOPRZYLEPNE TESA TACK 72 SZT. TRANSPARENTNE [59408-00000-00 TS]	4042448366849	1 kpl	20,00	7,99	7,99 P23	1,84		9,83

Rysunek 3.80: Fragment tabeli reprezentującej styl drugi ze zbioru testowego

Lp.	Column1	Symbol	Ilość	J.m.	Rabat %	Cena netto			
		Nazwa produktu			%	netto			
1	127321 FC	9556089009232			18 szt	20,00	2,15		
		GUMKA ARTYSTYCZNA CHLEBOWA MIX KOLETUPLASTIKOWE CZERWONA/NIEBIESKA/ZŁOTA FABER-CASTELL [127321 FC]							
2	300/001/C ED	4004764390564			10 szt	20,00	2,07		
		MARKER EDDING PERMANENTNY OKRĄGLA KONCÓWKA 1,5-3MM CZARNY [300/001/C ED]							
3	54410-21063 DA	4007885232898			10 szt		7,91		
		NOŻYCZKI DAHLE COMFORT-GRIP, DO PAPIERU 24,7 cm [54410-21063 DA]							
4	54405-21092 DA	4007885232867			10 szt		3,23		
		NOŻYCZKI DAHLE COMFORT-GRIP, DOMOWE 14 cm [54405-21092 DA]							
5	117201 FC	4005401172017			12 szt	20,00	2,15		
		OŁÓWEK GRIP 2001 B Z GUMKĄ FABER-CASTELLCASTELL [117201 FC]							
6	114000 FC	4005401140009			1 kpl	20,00	17,59		
		OŁÓWKI GOLD FABER 1222 6 SZT. +GUMKA+TEMPERÓWKA OPAKOWANIE KARTONOWE FABER-CASTELL [114000 FC]							
7	59408-00000-00 TS	4042448366849			1 kpl	20,00	7,99		
		PŁATKI SAMOPRZYLEPNE TESA TACK 72 SZT. TRANSPARENTNE [59408-00000-00 TS]							

Rysunek 3.81: Wyniki działania Camelot dla fragmentu tabeli z rysunku 3.80

Metoda autorska, poza wspomnianym stylem drugim, radziła sobie stosunkowo dobrze dla wszystkich pozostałych typów faktur. Problemem była wysoka dewiacja (powyżej 0,1) wyników dla stylu dziewiątego (rys. 3.7). Szczegółowa analiza plików wynikowych pokazała, że w przypadku tego konkretnego stylu, przy parametrach dobranych przez autora, ProjectionP myliła się czasami przy segmentacji kolumnowej (rys. 3.82).

Lp.	Kod Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
25	131 Kal.INFO A5 (mix)	5905824400904	5 szt.	5,7900	28,95	23%
26	132 Kal.INFO A6 (mix)	5905824600175	5 szt.	4,9800	24,90	23%
27	550 Kalendarz menadżera A5	5905824900303	5 szt.	7,5900	37,95	23%
28	126 Kalendarz nauczyciela A5	5905824400928	10 szt.	5,8800	58,80	23%
29	531 Kalendarz ucznia A6	5905824100668	20 szt.	2,5100	50,20	23%
30	175 Karton A1-20 kremowy	5905824900381	1 opak.	19,0000	19,00	23%
31	178 Karton A1-20 pomarańczowy	5905824800933	1 opak.	19,0000	19,00	23%
32	210 Kostka kolor 8,5x8,5 mix klejona	5905824000586	4 szt.	1,5600	6,24	23%
33	594 Mapa Europy -szkol.podkl.na biurko-intro	5905824400799	1 szt.	5,1500	5,15	5%
34	595 Mapa Polski -szkol.podkl.na biurko-intro	5905824200047	1 szt.	4,1300	4,13	5%
35	599 Mapa Świata -szkol.podkl.na biurko-intro	5905824400782	1 szt.	5,1500	5,15	5%
36	247 Papier (jk) A4-100 j.zielony	5905824801190	1 opak.	4,7500	4,75	23%
37	297 Papier (jk) A4-100 żółty (fluo)	5905824300839	2 opak.	4,9700	9,94	23%
38	492 Papier 150g A4-100 kredowany	5905824600069	1 opak.	6,7400	6,74	23%
39	269 Papier fluo mix A4-100 kolorowy	5905824500529	5 opak.	5,0900	25,45	23%
40	229 Papier mix A4-100 kolorowy	5905824200993	5 opak.	4,5300	22,65	23%
41	338 Papier mix A4-200 kolorowy 10 kolor.	5905824000814	2 opak.	8,9000	17,80	23%
42	233 Papier mix A4-250 kolorowy	5905824200986	1 opak.	10,4500	10,45	23%
43	2325 Szkicownik A4-90 90g na spirali ART expert	5905824500970	2 szt.	12,1500	24,30	23%
44	592 Szkicownik artystyczny A4	5905824200023	5 szt.	5,7900	28,95	23%
45	646 Teczka (jk) A3/10 szkolna z gumką lakier.	5905824130313	1 opak.	25,0000	25,00	23%
46	649 Teczka (jk) A4/10 szkolna z gumką lakier.	5905824110971	5 opak.	11,0000	55,00	23%
47	648 Teczka (jk) A4/10 szk.pastel.z gumką lakier	5905824010431	5 opak.	11,0000	55,00	23%
48	671 Teczka (TFK) A3/5 szkolna z gumką	5905824705795	3 opak.	21,8800	65,64	23%

Rysunek 3.82: Wyniki działania ProjectionP dla tabeli stylu dziewiątego z błędą segmentacją kolumnową

W przypadku, gdy przestrzeń pomiędzy kolumnami była mała i większość tekstu w kolumnie dobiegała do skraju swojej komórki, zdarzało się, że filtrowanie wykonane przez metodę autorską zamknęło doliny, przez co algorytm nie był w stanie wykryć miejsca segmentacji i ze spalał te kolumny. Dodatkowa manipulacja parametrami (zwiększenie wykładnika potęgowania i zmniejszenie długości filtra) była w stanie skorygować te błędy (rys. 3.83). Nie zmienia to jednak faktu, że problemy zostały zdiagnozowane dopiero po wykonaniu testów. Z nowymi parametrami metoda wydała się być stabilna - poprawiła się z 0.840 do 0.973 f1-score (w ramach tego stylu), jedynie błędy jakie popełniała to przecięcie słów „Kod” i „Vat” w nagłówkach. Nie można niesłety wykluczyć, że dodanie dodatkowych, trudnych przykładów do zbioru testowego mogłoby ponownie wyjawić niepożądane zachowania metody.

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

Lp.	Kod	Nazwa towaru	Kod EAN	Ilość/J.m.	Cena netto	Wartość netto	VAT
1	434	Blok akwarelowy Art Expert A4-10 300g	5905824101139	10 szt.	5,4900	54,90	23%
2	447	Blok milimetrowy A4-20	5905824600878	20 szt.	1,1100	22,20	23%
3	43	Blok Superior techniczny kolorowy A3-25	5905824300105	10 szt.	6,1300	61,30	23%
4	62	Blok Superior techniczny kolorowy A5-25	5905824900473	10 szt.	1,7500	17,50	23%
5	41	Blok tech. A3-10 250g	5905824801664	50 szt.	2,5200	126,00	23%
6	42	Blok tech. A4-10 250g	5905824801671	100 szt.	1,2900	129,00	23%
7	13	Blok techn. z kolorowymi kartkami A3-10	5905824000265	20 szt.	2,3000	46,00	23%
8	11	Blok techniczny A3-10	5905824000029	20 szt.	1,8000	36,00	23%
9	412	Blok techniczny kolorowy ELITE A4-30	5905824501281	10 szt.	3,6600	36,60	23%
10	91	Brulion 200k. A5 tw.opr.	5905824000289	5 szt.	5,2500	26,25	23%
11	93	Brulion 300k. A4 tw.opr.	5905824000432	2 szt.	12,8200	25,64	23%
12	106	Brulion 64k. A7 tw.opr.	5905824200252	10 szt.	1,1700	11,70	23%
13	177	Karton A1 niebieski (jasny)	5905824800797	1 opak.	19,0000	19,00	23%
14	749	Karton A3-20 atramentowy	5905824500222	1 opak.	4,9400	4,94	23%
15	784	Karton A3-20 bordowy	5905824500321	2 opak.	4,9400	9,88	23%
16	812	Karton A3-20 brązowy	5905824500338	2 opak.	4,9400	9,88	23%
17	813	Karton A3-20 brązowy (jasny)	5905824500208	2 opak.	4,9400	9,88	23%
18	814	Karton A3-20 cytrynowy	5905824500215	2 opak.	4,9400	9,88	23%
19	864	Karton A3-20 czarny	5905824200740	2 opak.	5,3400	10,68	23%
20	815	Karton A3-20 czerwony	5905824500314	2 opak.	4,9400	9,88	23%
21	816	Karton A3-20 czerwony (fluo)	5905824000951	1 opak.	4,9400	4,94	23%
22	817	Karton A3-20 fioletowy	5905824500345	2 opak.	4,9400	9,88	23%
23	843	Karton A3-20 granatowy	5905824500352	2 opak.	4,9400	9,88	23%
24	845	Karton A3-20 kreślony	5905824500239	2 opak.	4,9400	9,88	23%

Rysunek 3.83: Wyniki działania ProjectionP dla tabeli z rysunku 3.82 po korekcji parametrów

Analizując tabele stylu dziesiątego, zdiagnozowano jeszcze jedną niedoskonałość. W przypadku wystąpienia znaków poza właściwym obszarem tabeli, ProjectionP uwzględnia te znaki jako część tabeli i zapisuje je do pliku, często pocięte przez segmentację wykonaną przez algorytm. Przykładowo, na rysunku 3.84, na poziomie tabelki podsumowującej znajdują się znaki, które w prawdzie nie powodują błędów segmentacji, ale po przetworzeniu powodują zanieczyszczenia w pliku wyjściowym (ciągi nic nie znaczących znaków, rys. 3.85). Teoretycznie nie jest to problem samego algorytmu rozpoznawania struktury, a raczej detekcji obszaru tabeli. Komplikacja ta pokazuje również, że sprowadzenie problemu detekcji do znalezienia prostokątnego obszaru na obrazie nie musi być wystarczające dla każdego rodzaju dokumentu wejściowego. Niemniej jednak, dla praktycznego zaaplikowania wyników, autorska metoda powinna zostać wzbogacona o dodatkowe mechanizmy usuwania tych artefaktów w przetwarzaniu wstępny bądź końcowym.

ROZDZIAŁ 3. PRZEPROWADZONE EKSPERYMENTY I BADANIA

LP	Symbol towaru indeks odbiorcy/EAN	Nazwa towaru / (PKWiU)	Ilość J.M.	Cena jedn. bez podatku netto	Rabat %	Cena jedn. po rabacie	Wartość towaru/usł. po rabacie
1	VLP-4100118 5901478995032	PAPIER TOALET. VELVET, BIAŁY, 2W CELUL., OP. 8 ROL. PO 138 LISTKÓW (15 METRÓW) (17.22.11.0)	8 OP	3,47	0,00	3,47	27,76
2	3M-UU00487408 5 5902658076305	TAŚMA MONTAŻOWA SCOTCH, 19MMx1,5M, PRZEZROCZYSTA (22.29.21.0)	1 szt	11,80	0,00	11,80	11,80
3	3M-UU00638003 2 5900323008606	17001 PL HAKI COMMAND WIELOKROTNEGO UZYTKU, ŚREDNIE, 2 HAKI, 4 ŚREDNIE PASKI (22.29.22.0)	1 op	7,13	0,00	7,13	7,13
4	3M-UU00638031 3 5900323008385	17006 PL HACZYKI MAŁE, COMMAND, GSZT, 8 MINI PASKÓW (22.29.22.0)	1 szt	7,13	0,00	7,13	7,13
5	SKY-88031879 7318761070006	PAPIER KSERO SKY COPY, A4, 500 ARKUSZY, KLASA C, GRAMATURA 80 (17.12.14.0)	120 op	8,67	0,00	8,67	1 040,40
RAZEM							1 094,22
							1 094,22

Do zapłaty: 1 094,22

Rysunek 3.84: Fragment tabeli reprezentującej styl dziesiąty ze zbioru testowego

LP	Symbol towaru ind	Nazwa towaru / (PKWiU)	Ilość J.M.	Cena j	Rabat 0/0	Cena je	Wartość
1	VLF-41001 18 5901478 5 591	PAPIER TOALET. VELVET, BIAŁY, 2w CELUL., c 8 OP	3,47	0	3,47	27,76	
2	3M-UU00487408 5 591	TAŚMA MONTAŻOWA SCOTCH, 19MMX1,5M 1 szt	11,8	0	11,80	11,8	
3	3M-UU00638003 2 591	17001 PL HAKI COMMAND WIELOKROTNEGC 1 OD	7,13	0	7,13	7,13	
4	3M-UU00638031 3 591	17006 PL HACZYKI MAŁE, COMMAND, GSZT, t 1 szt	7,13	0	7,13	7,13	
5	SKY-88031879 731876	PAPIER KSERO SKY COPY, A4, 500 ARKUSZY, K 120 Op	8,67	0	8,67	1040,4	
							RAZEM
							1094,22
		>_l,_l_____"A'"A					
		I... II II.-nl..- . _1_-;-'_n'					

Rysunek 3.85: Wyniki działania ProjectionP dla fragmentu tabeli z rysunku 3.84 z artefaktami spowodowanymi obecnością tekstu poza obszarem tabeli

Stwierdzono wcześniej, że Nanonets nie popełniało błędów przy segmentacji kratowej, jednak nie było to do końca prawdą. Rzeczywiście, segmentacja obszarów na obrazie wizualnie wygłaąda niezawodnie (rys. 3.87), jednak po zastosowaniu modułu OCR i wygenerowaniu pliku wyjściowego, w niektórych przypadkach można było zaobserwować niepożądane zachowanie. W miejscach, gdzie tekst był bardzo bliski punktu segmentacji, moduł OCR czasami umieszczał słowo w niewłaściwej komórce lub równocześnie w dwóch komórkach (rys. 3.88). Prawdopodobnie było to rozwiązanie mające na celu zapobieganie przecinaniu słów przez moduł segmentujący. Nie zmieniało to faktu, że traktowano to jako błąd segmentacji w ramach badań ewaluacyjnych.

lp.	mag	kod kreskowy	nazwa towaru/usługi
1	A01	5902277171221	BLOK BIUROWY A4 50# 70G
2	A01	5902277171238	BLOK BIUROWY A5 50# 70G
3	A01	5902277070005	BLOK TECHNICZNY A4 10 170G
4	A01	5902277171269	KOŁOZESZYT A4 80# M 70G Z PERF.UV
5	A01	5902277213044	TECZKA Z GUMKĄ A4+ CZERWONA
6	A01	5902277213082	TECZKA Z GUMKĄ A4+ POMARAŃCZOWA

Rysunek 3.86: Fragment tabeli reprezentującej styl siódmy ze zbioru testowego

lp.	mag	kod kreskowy	nazwa towaru/usługi
1	A01	5902277171221	BLOK BIUROWY A4 50# 70G
2	A01	5902277171238	BLOK BIUROWY A5 50# 70G
3	A01	5902277070005	BLOK TECHNICZNY A4 10 170G
4	A01	5902277171269	KOŁOZESZYT A4 80# M 70G Z PERF.UV
5	A01	5902277213044	TECZKA Z GUMKĄ A4+ CZERWONA
6	A01	5902277213082	TECZKA Z GUMKĄ A4+ POMARAŃCZOWA

Rysunek 3.87: Wyniki segmentacji metody Nanonets na fragmencie tabeli z rysunku 3.86

lp.	mag	kodkreskowy	nazwatowaruu/usługi
1	A01	5902277171221 BLOK	BIUROWY A4 50# 70G
2	A01	5902277171238 BLOK	BIUROWY A5 50# 70G
3	A01	A01 5902277070005	BLOK TECHNICZNY A4 10 170G
4	A01	5902277171269	KOŁOZESZYT A4 80# M 70G Z PERF.UV
5	A01	5902277213044 TECZKA Z GUMKĄ A4+ CZERWONA	
6	6 A01	5902277213082	TECZKA Z GUMKĄ A4+ POMARAŃCZOWA

Rysunek 3.88: Końcowe wyniki działania metody Nanonets na fragmencie tabeli z rysunku 3.86

Wyżej opisany problem wpływał na to, że metoda ProjectionP miała lepsze rezultaty niż Nanonets przy rozpoznawaniu struktury stylów siódmego i dziesiątego. Fakt, że każda z metod miała zestaw stylów, w którym się specjalizowała zainspirował wykonanie jeszcze jednej analizy.

Biblioteka Camelot jest rozwiązaniem open - source, przez co teoretycznie możliwe jest zintegrowanie jej z ProjectionP, przykładowo jako dodatkowego trybu przetwarzania. Porównano

więc działanie rozwiązania hybrydowego złożonego z ProjectionP i Camelot z rozwiązaniem komercyjnym Nanonets. Dla każdego stylu tabeli wzięto wyniki należące do lepszej z dwóch metod, następnie ponownie wyznaczono wyniki zbiorowe (tabela 3.7).

Po połączeniu najlepszych wyników, nowo powstała w ten sposób metoda działała lepiej i stabilniej od Nanonets. W istocie potrafiła ona poradzić sobie w zadowalającym stopniu z wszystkimi stylami tabel należącymi do zbioru, również z tymi mocno nieoczywistymi, jak styl drugi i piąty.

	Camelot or ProjectionP*		Nanonets	
	ICDAR2013 eval.	Modified* eval.	ICDAR2013 eval.	Modified* eval.
avg_f1-score	0.9552	0.9870	0,9432	0,9785
std_f1-score	0.0751	0.0219	0,1141	0,0459

Tablica 3.7: Zestawienie wyników połączonych metod Camelot i ProjectionP z metodą Nanonets

4. Podsumowanie

4.1 Spostrzeżenia i wnioski

Przyjęte przez autora podejście algorytmiczne okazało się być bardzo skuteczne w kontekście zaproponowanego zbioru danych. W połączeniu z pozostałymi wynikami osiągniętymi przez zestawiane metody, można stwierdzić, że teza niniejszej pracy, zakładająca, że możliwe jest opracowanie metod rozpoznawania tabel osiągających bardzo dobre, predysponujące do praktycznego wykorzystania wyniki została udowodniona.

Sam opracowany algorytm prezentuje się bardzo dobrze dla określonych stylów, a możliwość manipulacji dodatkowymi parametrami pozwala często dostroić go do stopnia, w którym nie popełnia istotnych błędów dla tabel prostych i półprostych (zdarza się, że przecina słowa w nagłówku, co da się wyeliminować w post - processingu).

Niestety, w praktyce można napotkać na bardzo szeroką gamę różnych stylów tabel i pewnym jest, że zastosowane proste podejście decyzji progowych nie będzie w stanie poradzić sobie z każdym z nich (przykładowo, nie zawsze w tabelach nieobramowanych tekst jest wyrównywany do góry). Nie jest to jednak mankamentem jedynie metody autorskiej. W literaturze można znaleźć liczne przykłady, w których autorzy pokazują typy tabel, z których proponowane rozwiązania nie są w stanie poprawnie przeanalizować [27], [25], [36], [1], [29]. Wykonane analizy pokazały, że podobne problemy dotyczą również testowanych Camelot i Nanonets.

Innym istotnym zagadnieniem jest zasadność praktycznego wykorzystania silnie parametryzowanej metody, jaką jest ProjectionP. Bez wątpienia ma ona sens, w przypadku, gdy wtórnie analizowany jest pewien określony zestaw stylów tabel nie różniących się znacząco w obrębie danej klasy. Taka sytuacja, według wiedzy autora, zdarza się często w prywatnych przedsiębiorstwach, w których na przestrzeni miesięcy lub lat prowadzone są interesy z tymi samymi kontrahentami, a jedynie od czasu do czasu zmieniają się typy dokumentów wymagające analizy. Z drugiej strony, autorska metoda może nie być szczególnie przydatna dla kogoś, kto zajmuje się pozyskiwaniem dużej ilości informacji z różnych źródeł (np. osobie analizującej dokumenty archiwalne różnej genezy w celu wydobycia danych do analiz i prognozowania). Taka osoba z pewnością wielokrotnie musiałaby dobierać parametry dostosowane do stylu dokumentu, przez co cała procedura mogłaby

być nieopłacalna w porównaniu z innymi rozwiązaniami.

Dokonane zestawienie metody opartej na parsowaniu pliku PDF (Camelot) z algorytmami analizującymi obraz (ProjectionP, Nanonets) potwierdziło kilka spostrzeżeń często pojawiających się w literaturze. Z pewnością metadane pozwalają bezbłędnie wydobyć tekst z dokumentu, co w dużym stopniu ułatwia również jego klasteryzację do poziomu komórek (przykładowo przy analizie drugiego stylu tabeli, rys. 3.80) i dzięki temu dając algorytmom istotną przewagę. Istnieją jednak typy wizualnych przesłanek świadczących o segmentacji (jak obramowania zwykłe i przerywane, geometryczne ułożenie tekstu), które w bardziej naturalny sposób identyfikuje się przy pomocy algorytmów widzenia komputerowego, przez co dodatkowa analiza plików PDF wygenerowanych cyfrowo po przekonwertowaniu ich do postaci obrazów ma swoje zastosowanie.

Wnikając głębiej w techniczne szczegóły zaproponowanego algorytmu, zaobserwowano, że metoda oparta jedynie na projekcji profilowej obszaru tabeli jest zbyt ograniczona, by mogła poradzić sobie z dokładną analizą złożonych struktur. Może natomiast być stosowana z powodzeniem, jako algorytm podstawowy lub wspomagający rozpoznawanie kraty. Ciekawym pomysłem wydaje się być łączenie wyników kilku różnych podejść do segmentacji, co zostało kilkukrotnie zastosowane w ramach ProjectionP. Należy pamiętać jednak, że takie łączenie może sumować również błędy obu metod, a w niektórych miejscach powodować powstanie pustych kolumn / wierszy generowanych przez wielokrotną, prawidłową segmentację, nie wychwyconą przez działania post-processingowe.

Możliwym wydaje się być doprowadzenie opracowanego algorytmu do wersji ze zmniejszoną liczbą dodatkowych parametrów lub całkowicie autonomicznej. Można to zrealizować poprzez proste estymowanie tych parametrów z wykorzystaniem pobranych z obrazu cech na podobieństwo metody Zuyeva [36] lub wykorzystania modułów sztucznej inteligencji z uczeniem nienadzorowanym [35]. Niestety, wielce prawdopodobnym jest, że zwiększenie stopnia automatyzacji poskutkuje również istotnym pogorszeniem dokładności systemu.

W celu podjęcia prób całkowitej automatyzacji, ciekawszym pomysłem wydaje się być zmiana podejścia na system end - to - end z wykorzystaniem konwolucyjnych sieci neuronowych i spostrzeżeń wynikających z opracowanej metody top-down, tak by równolegle obejmowana była również detekcja tabeli. Za podstawę systemu mogłyby posłużyć rozwiązanie TableNet [25], którego autorzy chwalą się dobrymi wynikami dla segmentacji kolumnowej, ale nie prezentują żadnego zaawansowanego modułu segmentacji wierszowej. Moduł oparty na klasyfikacji linii tekstu zaimplementowany w ramach ProjectionP radził sobie dobrze, nawet z nieoczywistymi przykładami tabel, dlatego zimportowanie go do rozwiązania TableNet i dodatkowe dopracowanie go (np. przez dodanie dodatkowych cech wykorzystywanych przy klasyfikacji) mogłyby, zdaniem autora, przynieść dobre końcowe efekty.

Sama metodyka rozwijania algorytmu zastosowana przy realizacji pracy doprowadziła osta-

tecznie do osiągnięcia pozytycznych rezultatów, ale wiązała się z wieloma problemami i ograniczeniami. Wizualna ocena wyników przetwarzania sygnałów i rezultatów końcowych pozwalała na wyciągnięcie wnikliwych informacji o wprowadzanych modyfikacjach. Niestety, w kluczowych momentach potrzebna była ewaluacja wprowadzonych zmian na całym zbiorze treningowym i niejednokrotnie okazywało się, że wprowadzone modyfikacje skutkowały poprawieniem wyników tylko dla części tabel ze zbioru, pogarszając tym samym wyniki dla innych. Wymuszało to wprowadzanie szeregu poprawek i powtórzenie procesu ewaluacji lub zrezygnowaniem z wprowadzonych zmian.

Oprócz tego, jak pokazano w rozdziale 3.6, ostatecznie w fazie testów wychwycono błędy, których nie zaobserwowano na zbiorze treningowym. Świadczy to o fakcie, że objętość zbioru treningowego nie była wystarczająca i powinno się w nim znaleźć więcej wymagających przykładów. Z drugiej strony, zwiększenie liczby przykładów w zbiorze wiązałoby się z wydłużeniem i skomplikowaniem procesu rozwijania algorytmu.

Niezbędnym wydaje się być dołączenie do procesu oceny rezultatów, automatycznej metody ewaluacji, która w dobrym stopniu oddawałaby skuteczność działania algorytmu uwzględniając tym samym specyfikę wyników, jakie mają zostać osiągnięte. Pozostaje to problemem otwartym, z którym w dalszym ciągu borykają się autorzy publikacji [18], [7], [35],[8].

Zastosowane metryki wykorzystane przy ewaluacji końcowej, jak i sam sposób ich wyliczania z wykorzystaniem sporzązonego narzędzia zdaniem autora sprawdziły się dobrze dla zaproponowanego zbioru danych i zestawianych algorytmów. Końcowe rezultaty dość dobrze oddają to, jaka jest wizualna ocena jakości poszczególnych wynikowych plików CSV. Należy jednak zaznaczyć, że sam proces obsługi narzędzia jest dość pracochłonny i wymaga od użytkownika nabrania pewnej wprawy.

Wprowadzona modyfikacja do często cytowanej implementacji ewaluacji ICDAR2013 również spełniła postawione jej założenia. Była w stanie lepiej podkreślić różnicę pomiędzy precyzją, a czułością i w mniejszej restrykcyjny sposób oceniała tabele tylko częściowo poprawnie przeanalizowane.

To, czy modyfikacja metody ewaluacji jest bardziej użyteczna od oryginału wykorzystywanego w konkursie ICDAR2013, pozostaje kwestią dyskusyjną. Z pewnością, jeśli priorytetem jest silne penalizowanie błędów segmentacji (często sprawiających, że wynik końcowy jest bezużyteczny) oryginał sprawdzi się lepiej. W przypadku, gdy celem jest sprawdzenie, czy algorytm ma większe problemy z nadmierną czy niewystarczającą segmentacją, lepszą oceną wydaje się być modyfikacja.

W obu przypadkach podstawowym problemem pozostaje automatyzacja. Zdaniem autora obie metody ewaluacji mogą zostać zautomatyzowane z wykorzystaniem zaawansowanych metod porównywania tekstu takich jak miara BLEU [7] (w sposób bardziej sprawiedliwy niż implemen-

tacja konkursowa sprawdzająca zgodność znaków [16]). Wymaga to jednak dużego nakładu pracy. Dobrym początkiem wydaje się być wyjście od plików wynikowych przygotowanych w ramach wykonanych prac. Wyniki ewaluacji zostały przygotowane z wnikliwą, ludzką interpretacją, więc mogą posłużyć jako ground - truth w próbach automatyzacji.

4.2 Przyszłe prace

Nie ulega wątpliwości, że temat, którego dotyczy niniejsza praca jest mocno rozbudowany i osiągnięte efekty tylko w niewielkim stopniu rozwiązują ogólny problem automatycznego rozpoznawania tabeli w dokumentach. Ciężko jest odmówić zagadnieniu praktycznego zastosowania, czego potwierdzeniem jest rosnące zainteresowanie tematem w środowisku komercyjnym i badawczym [18].

Rozważane w pracy problemy i opracowane dotychczas rozwiązania pozwalają na wskazanie kierunków dalszych prac:

- Dostosowanie opracowanego algorytmu do założeń publicznie dostępnego zbioru danych wykorzystanego w konkursie ICDAR2013 [16], przetestowanie go i dalszy rozwój w oparciu o poczynione spostrzeżenia
- Przetestowanie opracowanego algorytmu na skanach dokumentów i dostrojenie go do plików gorszej jakości
- Opracowanie rozwiązania end - to - end opartego na konwolucyjnych sieciach neuronowych, wspieranego już uzyskanymi efektami
- Podjęcie prac nad automatyzacją półautomatycznej implementacji metod oceniania algorytmów wykorzystanej w ramach niniejszej pracy

Rozpoczęte badania, będą przez autora pracy kontynuowane. W pierwszej kolejności prawdopodobnie wypróbowana zostanie zmiana podejścia na model wykorzystujący głębokie uczenie, gdyż może ona ukazać dodatkową perspektywę w kontekście możliwych do wykorzystania rozwiązań.

Generalnie, można stwierdzić, że prace zakończyły się w sporej części sukcesem. Nie udało się w prawdzie zautomatyzować w stu procentach procesu rozpoznawania tabeli (od detekcji po wygenerowanie pliku wyjściowego), ale udało się doprowadzić osiągane wyniki do dużego stopnia niezawodności na pełnym przekroju branych pod uwagę danych wejściowych. Rezultaty prac na pewno można wdrożyć w formie wyspecjalizowanej aplikacji wspomagającej prowadzenie rozliczania faktur w przedsiębiorstwach. Ostatecznie, przy tak wrażliwym problemie jak obsługa systemu księgowego, niezawodność zawsze ma pierwszeństwo przed wygodą.

Spis tabelic

2.1	Zestawienie typów metod rozpoznawania tabel sporządzone przez autora na podstawie wniosków z zebranej literatury	13
2.2	Zestawienie wybranych metod z literatury pod kątem wyników detekcji tabel	14
2.3	Zestawienie wybranych wyników z literatury pod kątem rozpoznawania struktury tabeli	21
3.1	Liczebności tabel w zaproponowanej bazie ze względu na typ obramowania i złożoność	26
3.2	Liczebności tabel w zbiorze treningowym ze względu na typ obramowania i złożoność	26
3.3	Liczebności tabel w zbiorze testowym ze względu na typ obramowania i złożoność	26
3.4	Statystyki dotyczące liczby kolumn i wierszy tabel w zbiorze treningowym	26
3.5	Statystyki dotyczące liczby kolumn i wierszy tabel w zbiorze testowym	26
3.6	Wyniki osiągnięte na pełnym zbiorze testowym przez zestawiane metody	77
3.7	Zestawienie wyników połączonych metod Camelot i ProjectionP z metodą Nanonets	85

Spis rysunków

2.1	Przykład tabeli o złożonym układzie, rysunek pochodzący z publikacji Zanibbiego [34]	9
2.2	Opis układu tabeli z rysunku 2.1 w odniesieniu do jej kraty (wiersze kraty są indeksowane od 1 do 6, kolumny od A do D), rysunek pochodzący z publikacji Zanibbiego [34]	9
2.3	Wyszczególnienie regionów tabeli, rysunek pochodzący z publikacji Zanibbiego [34] i zmodyfikowany przez autora	9
2.4	Przykład tabeli prostej pochodzącej z autorskiego zbioru danych	10
2.5	Przykład tabeli złożonej pochodzącej z autorskiego zbioru danych	11
2.6	Szablon pliku wynikowego branego pod uwagę w konkursie ICDAR - 2013	16
2.7	Szablon pliku wynikowego branego pod uwagę w konkursie ICDAR - 2019	16
2.8	Tabela o złożonej strukturze	17
2.9	Konwersja tabeli (rys. 2.8) o złożonej strukturze do tabeli prostej wykorzystywana przez Camelot	17
2.10	Zasady oceniania połączeń w tabelach, rysunek pochodzący z pracy Gobela [14] .	18
3.1	Tabela F-0034-10-G-21_0 pochodząca z autorskiego zbioru danych	27
3.2	Tabela 2021_FA_MG01_2285_1 pochodząca z autorskiego zbioru danych	27
3.3	Tabela dok_int2_0 pochodząca z autorskiego zbioru danych	28
3.4	Tabela FSK_1803_0807_1 pochodzącej z autorskiego zbioru danych	29
3.5	Tabela FV210310097_0 pochodząca z autorskiego zbioru danych	29
3.6	Tabela gd7.jrUT9w3LT_FV35021H2A2021_1 pochodząca z autorskiego zbioru danych	30
3.7	Tabela kr6_1 pochodząca z autorskiego zbioru danych	30
3.8	Wyniki detekcji komórek osiągnięte przez niemodyfikowany model CascadeTabNet na tabeli F-0034-10-G-21_0 ze zbioru treningowego	31
3.9	Wyniki detekcji komórek osiągnięte przez niemodyfikowany model CascadeTabNet na tabeli FSK_1803_0807_1 ze zbioru treningowego	32

3.10 Wyniki detekcji komórek osiągnięte przez niemodyfikowany model CascadeTab-Net na tabeli 2020_FA_MG01_14053_0 ze zbioru treningowego	32
3.11 Wyniki detekcji słów osiągnięte przez moduł <i>pytesseract</i> na przykładowej tabeli F-0034-10-G-21_0 ze zbioru treningowego	34
3.12 Wyniki detekcji słów osiągnięte przez moduł <i>pytesseract</i> na tabeli FA_12-10-18-Fwz-HS_Oryginal_1_16_1 ze zbioru treningowego	35
3.13 Wyniki detekcji słów osiągnięte przez moduł <i>pytesseract</i> na tabeli 2020_FA_MG01_14053_0 ze zbioru treningowego	35
3.14 Tabela F-0034-10-G-21_0 ze zbioru treningowego po binaryzacji	36
3.15 Projekcja pikseli tabeli z rys. 3.14 na oś poziomą	37
3.16 Projekcja pikseli tabeli z rys. 3.14 na oś pionową	37
3.17 Tabela dok_int2_0 ze zbioru treningowego po binaryzacji	38
3.18 Projekcja pikseli tabeli z rys. 3.17 na oś poziomą	38
3.19 Projekcja pikseli tabeli z rys. 3.17 na oś pionową	39
3.20 Tabela kr1_0 ze zbioru treningowego po binaryzacji	39
3.21 Projekcja pikseli tabeli z rys. 3.20 na oś poziomą	40
3.22 Projekcja pikseli tabeli z rys. 3.20 na oś pionową	40
3.23 Interfejs graficzny przygotowanej aplikacji	44
3.24 Ogólny schemat działania opracowanej metody	45
3.25 Tryby segmentacji	46
3.26 Schemat działania detekcji obramowań	47
3.27 Tabela kr6_0 z przerywanymi liniami	48
3.28 Segmentacja tabeli z rysunku 3.27 w osi pionowej metody opartej o morfologię	49
3.29 Projekcja tabeli 3.27 na oś pionową	49
3.30 Projekcja tabeli 3.27 na oś pionową po transformacji (w stosunku do średniej)	50
3.31 Wyniki segmentacji tabeli 3.27 w osi pionowej przez metodę wykorzystującą projekcję	50
3.32 Tabela gd7e9pWKy0yL_FV35231H2A2021_0 z tabelką podsumowującą	51
3.33 Segmentacja tabeli z rysunku 3.32 w osi pionowej metody opartej o morfologię	51
3.34 Projekcja tabeli 3.32 na oś pionową	51
3.35 Projekcja tabeli 3.32 na oś pionową po transformacji (w stosunku do średniej)	52
3.36 Wyniki segmentacji tabeli 3.32 w osi pionowej przez metodę wykorzystującą pojęgowanie i projekcję	52
3.37 Schemat działania segmentacji bez obramowań	53
3.38 Tabela 2018_FA_MG01_24115_0 pionowo obramowana ze zbioru treningowego	55
3.39 Surowa projekcja tabeli z rysunku 3.38 na oś pionową	55

3.40 Sygnał z rysunku 3.39 przetransformowany i znormalizowany w zakresie wartości (0,1)	56
3.41 Tabela dok_interdruk2_0 pionowo obramowana ze zbioru treningowego	56
3.42 Surowa projekcja tabeli z rysunku 3.41 na oś pionową	57
3.43 Sygnał z rysunku 3.42 przetransformowany i znormalizowany w zakresie wartości (0,1)	57
3.44 Tabela FV210310097 poziomo obramowana ze zbioru treningowego	58
3.45 Surowa projekcja tabeli z rysunku 3.44 na oś poziomą	58
3.46 Sygnał z rysunku 3.45 przetransformowany i znormalizowany w zakresie wartości (0,1)	59
3.47 Tabela kr1_0 poziomo obramowana ze zbioru treningowego	59
3.48 Surowa projekcja tabeli z rysunku 3.47 na oś poziomą	60
3.49 Sygnał z rysunku 3.48 przetransformowany i znormalizowany w zakresie wartości (0,1)	60
3.50 Schemat wnioskowania segmentacji kolumnowej bez korzystania z obramowań . .	61
3.51 Przykład „ogona” kolumny, powstałego na skutek wyrównywania tekstu	61
3.52 Przykład błędne wnioskowania spowodowanego niewzględnieniem ogona . . .	62
3.53 Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sy- gnału z rysunku 3.46	63
3.54 Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sy- gnału z rysunku 3.49	63
3.55 Schemat wnioskowania segmentacji wierszowej bez korzystania z obramowań . .	64
3.56 Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sy- gnału z rysunku 3.40	65
3.57 Końcowe wyniki segmentacji osiągnięte poprzez wnioskowanie na podstawie sy- gnału z rysunku 3.43	65
3.58 tabela FV210310097_0 z poszarzonym wierszem ze zbioru treningowego	67
3.59 tabela z rysunku 3.58 po binaryzacji z progiem podstawowym	67
3.60 tabela z rysunku 3.58 po binaryzacji z progiem dodatkowym	68
3.61 tabela z rysunku 3.58 po kroku 2. ($bin_{dif} = bin_{p_5} - bin_{p_1}$) przetwarzania	68
3.62 tabela z rysunku 3.58 po kroku 3. (usuwanie zaszumień) przetwarzania	68
3.63 tabela z rysunku 3.58 po kroku 4. (wypełnianie luk) przetwarzania	69
3.64 tabela z rysunku 3.58 po kroku 5. (ekstrakcja konturów) przetwarzania	69
3.65 Wyniki ekstrakcji informacji z tła tabeli z rysunku 3.58	69
3.66 Wynikowy plik CSV dla tabeli z rys. 3.44	70
3.67 Wynikowy plik CSV dla tabeli z rys. 3.38	70

3.68 Wynikowy plik CSV dla tabeli z rys. 3.41	71
3.69 Wynikowy plik CSV dla tabeli z rys. 3.47	71
3.70 Fragment tabeli kr_1 ze zbioru treningowego	75
3.71 Ewaluacja metody autorskiej na fragmencie tabeli z rys. 3.70 przy pomocy metodologii ICDAR2013	75
3.72 Ewaluacja metody autorskiej na fragmencie tabeli z rys. 3.70 przy pomocy metodologii zmodyfikowanej	75
3.73 Fragment tabeli 2020_FA_MG01_14153_0 ze zbioru treningowego	76
3.74 Ewaluacja Camelot autorskiej na fragmencie tabeli z rys. 3.73 przy pomocy metodologii ICDAR2013	76
3.75 Ewaluacja metody Camelot na fragmencie tabeli z rys. 3.73 przy pomocy metodologii zmodyfikowanej	76
3.76 Zestawienie uśrednionego f1 - score obliczonego metodą ewaluacji ICDAR2013 z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)	78
3.77 Zestawienie odchylenia standardowego f1 - score obliczonego metodą ewaluacji ICDAR2013 z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego) . .	78
3.78 Zestawienie uśrednionego f1 - score obliczonego zmodyfikowaną metodą ewaluacji z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)	79
3.79 Zestawienie odchylenia standardowego f1 - score obliczonego zmodyfikowaną metodą ewaluacji z uwzględnieniem stylu tabeli (właściciela dokumentu źródłowego)	79
3.80 Fragment tabeli reprezentującej styl drugi ze zbioru testowego	80
3.81 Wyniki działania Camelot dla fragmentu tabeli z rysunku 3.80	80
3.82 Wyniki działania ProjectionP dla tabeli stylu dziewiątego z błędnią segmentacją kolumnową	81
3.83 Wyniki działania ProjectionP dla tabeli z rysunku 3.82 po korekcji parametrów . .	82
3.84 Fragment tabeli reprezentującej styl dziesiąty ze zbioru testowego	83
3.85 Wyniki działania ProjectionP dla fragmentu tabeli z rysunku 3.84 z artefaktami spowodowanymi obecnością tekstu poza obszarem tabeli	83
3.86 Fragment tabeli reprezentującej styl siódmy ze zbioru testowego	84
3.87 Wyniki segmentacji metody Nanonets na fragmencie tabeli z rysunku 3.86	84
3.88 Końcowe wyniki działania metody Nanonets na fragmencie tabeli z rysunku 3.86 .	84

Bibliografia

- [1] Soumi Bardhan. “Table Detection and Text Extraction - OpenCV and Pytesseract”. W: *Analytics Vidhya* (2020).
- [2] CamelotDevelopers. *Camelot: PDF Table Extraction for Humans*. 2021. URL: <https://camelot-py.readthedocs.io/en/master/>.
- [3] Bidyut. B. Chaudhuri. “Digital document processing : major directions and recent advances”. W: 2006.
- [4] Stéphane Clinchant i in. “Comparing Machine Learning Approaches for Table Recognition in Historical Register Books”. W: *13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018*. IEEE Computer Society, 2018, s. 133–138. DOI: 10.1109/DAS.2018.44.
- [5] Bertrand Coüasnon i Aurélie Lemaitre. “Recognition of Tables and Forms”. W: *Handbook of Document Image Processing and Recognition*. Red. David S. Doermann i Karl Tombre. Springer, 2014, s. 647–677. DOI: 10.1007/978-0-85729-859-1_20. URL: https://doi.org/10.1007/978-0-85729-859-1_20.
- [6] Soumya De. *Table Extraction using Deep Learning*. Czer. 2021. URL: <https://medium.com/analytics-vidhya/table-extraction-using-deep-learning-3c91790aa200>.
- [7] Yuntian Deng, David Rosenberg i Gideon Mann. “Challenges in End-to-End Neural Scientific Table Recognition”. W: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, s. 894–901. DOI: 10.1109/ICDAR.2019.00148.
- [8] Hervé Déjean i in. *ICDAR 2019 Competition on Table Detection and Recognition (cTDeR)*. <http://sac.founderit.com/>. Kw. 2019. DOI: 10.5281/zenodo.3239032. URL: <https://doi.org/10.5281/zenodo.3239032>.
- [9] Tapio Elomaa. “ANSSI NURMINEN ALGORITHMIC EXTRACTION OF DATA IN TABLES IN PDF DOCUMENTS”. W: 2013.
- [10] Jing Fang i in. *A Table Detection Method for Multipage PDF Documents via Visual Separators and Tabular Structures*. 2011. DOI: 10.1109/icdar.2011.304.

- [11] Jing Fang i in. “Table Header Detection and Classification”. W: *Proceedings of the AAAI Conference on Artificial Intelligence* (2012).
- [12] Azka Gilani i in. “Table Detection Using Deep Learning”. W: Kyoto, Japan. T. 01. Kyoto, Japan: IEEE, 2017, s. 771–776. ISBN: 978-1-5386-3587-2. DOI: 10.1109/ICDAR.2017.131.
- [13] Hjalmar Gislason. *Data tables: From Sumer to VisiCalc*. Lip. 2016. URL: <https://hjallimedium.com/data-tables-from-sumer-to-visicalc-7b4d7b5a2150>.
- [14] Max C. Göbel i in. “A methodology for evaluating algorithms for table understanding in PDF documents”. W: *ACM Symposium on Document Engineering, DocEng ’12, Paris, France, September 4-7, 2012*. Red. Cyril Concolato i Patrick Schmitz. ACM, 2012, s. 45–48. DOI: 10.1145/2361354.2361365.
- [15] Max Göbel i in. “ICDAR 2013 Table Competition”. W: *2013 12th International Conference on Document Analysis and Recognition*. 2013, s. 1449–1453. DOI: 10.1109/ICDAR.2013.292.
- [16] Max Göbel i in. *ICDAR 2013 Table Competition*. 2013. URL: <https://roundtrippdf.com/en/data-extraction/dataset-format/>.
- [17] John Handley. “Table analysis for multi-line cell identification”. W: (sty. 2001). DOI: 10.1117/12.410853.
- [18] Khurram Azeem Hashmi i in. “Current Status and Performance Analysis of Table Recognition in Document Images With Deep Neural Networks”. W: *IEEE Access* 9 (2021), s. 87663–87685. DOI: 10.1109/ACCESS.2021.3087865.
- [19] Yilun Huang i in. “A YOLO-Based Table Detection Method”. W: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, s. 813–818. DOI: 10.1109/ICDAR.2019.00135.
- [20] Thong Huynh-Van i in. *Learning to detect tables in document images using line and text information*. 2018. DOI: 10.1145/3184066.3184091.
- [21] Katsuhiko Itonori. “Table structure recognition based on textblock arrangement and ruled line position”. W: *2nd International Conference Document Analysis and Recognition, ICDAR ’93, October 20-22, 1993, Tsukuba City, Japan*. IEEE Computer Society, 1993, s. 765–768. DOI: 10.1109/ICDAR.1993.395625.
- [22] Thomas Kieninger i Andreas Dengel. *The T-Recs Table Recognition and Analysis System*. 1999. DOI: 10.1007/3-540-48172-9_21.
- [23] Matthias Lee. *pytesseract 0.3.9*. Lut. 2022. URL: <https://pypi.org/project/pytesseract/>.

- [24] NanoNetTechnologiesInc. *Automated Table Extraction*. 2022. URL: <https://nanonets.com/table-extraction>.
- [25] Shubham Singh Paliwal i in. *TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images*. 2019. DOI: 10.1109/icdar.2019.00029.
- [26] Devashish Prasad i in. *CascadeTabNet - Github Repository*. Sierp. 2021. URL: <https://github.com/DevashishPrasad/CascadeTabNet>.
- [27] Devashish Prasad i in. *CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents*. 2020. DOI: 10.1109/cvprw50498.2020.00294.
- [28] Adrian Rosebrock. *Intersection over Union (IoU) for object detection*. List. 2016. URL: <https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/>.
- [29] Sebastian Schreiber i in. *DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images*. 2017. DOI: 10.1109/icdar.2017.192.
- [30] Tesseract-ocrTeam. *Tesseract documentation*. 2022. URL: <https://tesseract-ocr.github.io/>.
- [31] TheVisualCommunicationGuy. *A Brief History of How PDF Became Such a Popular File Format*. Kw. 2019. URL: <https://thevisualcommunicationguy.com/2019/04/15/a-brief-history-of-how-pdf-became-such-a-popular-file-format/>.
- [32] Prithiv S Vihar Kurama. *Table Detection, Table Extraction and Information Extraction using DL*. Czer. 2022. URL: <https://nanonets.com/blog/table-extraction-deep-learning/>.
- [33] Xinxin Wang. “Tabular Abstraction, Editing, and Formatting”. W: (wrz. 2000).
- [34] Richard Zanibbi, Dorothea Blostein i JamesR. Cordy. “A survey of table recognition. Models, observations, transformations, and inferences”. W: 7 (2004). ISSN: 1433-2833. DOI: 10.1007/s10032-004-0120-9.
- [35] Arthur Zucker i in. “ClusTi: Clustering Method for Table Structure Recognition in Scanned Images”. W: 26 (2021), s. 1765–1776. ISSN: 1383-469X. DOI: 10.1007/s11036-021-01759-9.
- [36] K. Zuyev. “Table image segmentation”. W: Ulm, Germany. T. 2. Ulm, Germany: IEEE, 1997, 705–708 vol.2. ISBN: 0-8186-7898-4. DOI: 10.1109/ICDAR.1997.620599.