

ДОСЛІДЖЕННЯ МЕТОДІВ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ ДОКУМЕНТІВ

Григор'єв Р.О

e-mail: rostyslav.hryhoriev@nure.ua

Харківський національний університет радіоелектроніки каф. ІМІ
м. Харків, Україна

This paper investigates the methods of document image segmentation with a specific focus on structural layout analysis. The study explores the application of hierarchical representations, namely M-X-Y trees, for the decomposition of document images into logical blocks (text, graphics, tables). The described method demonstrates high efficiency in establishing spatial relationships between segmented elements. The results confirm that utilizing M-X-Y trees significantly improves the accuracy of complex document reconstruction during the optical character recognition (OCR) process.

Стрімкий розвиток інформаційних технологій та необхідність масового оцифрування архівних фондів вимагають створення надійних систем оптичного розпізнавання символів (OCR). Однак перед безпосереднім розпізнаванням тексту система повинна «зрозуміти» структуру сторінки. Тому одним із найважливіших етапів попередньої обробки є структурна сегментація зображення документа – процес геометричної декомпозиції макета на окремі логічні блоки (абзаци, заголовки, ілюстрації) та встановлення порядку їх читання.

Метою даної роботи є дослідження методів структурної сегментації зображень документів та аналіз ефективності застосування ієрархічних деревоподібних структур для моделювання їх макетів.

Традиційні методи піксельної сегментації (бінаризація, виділення контурів) дозволяють відокремити корисний сигнал від фону, проте вони не зберігають інформацію про просторову топологію об'єктів. Для вирішення задачі аналізу складних багатоклонкових документів у дослідженні розглянуто алгоритми низхідного (top-down) розбиття сторінки, зокрема на основі методу рекурсивних X-Y розрізів.

Цей підхід базується на пошуку неперервних горизонтальних (X-cut) та вертикальних (Y-cut) пробілів (прогалін) між елементами документа. Процес розбиття макета зручно подавати у вигляді спеціалізованого графа – M-X-Y дерева. Для наочної демонстрації розробленої логіки структурної сегментації було побудовано відповідну модель на тестовому макеті (рис. 1).

Як видно з наведеного рисунка, на початкових етапах (фрагменти А, В, С) алгоритм виконує послідовне розбиття сторінки на вкладені прямокутні області. Кожному такому розрізу відповідає вузол у сформованому M-X-Y дереві (фрагмент D), де коренем є вся сторінка, а листками дерева виступають кінцеві сегментовані блоки (текстові чи графічні зони).

Приклад М-Х-У дерева

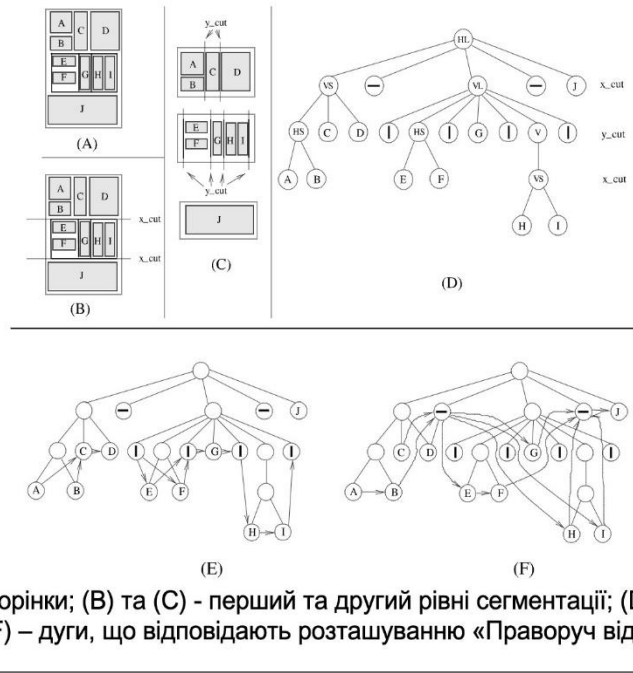


Рисунок 1 – Моделювання макета документа за допомогою М-Х-У дерева та графів просторових відношень

Окремою перевагою досліджуваного методу є можливість формалізації взаємного розташування елементів. Графи, подані на фрагментах (E) та (F), ілюструють автоматичне встановлення просторових відношень між виділеними сегментами, таких як «Праворуч від» та «Нижче від». Це критично важливо для правильного відновлення логічного порядку читання після сегментації.

Висновок: Проведене дослідження підтверджує, що для якісної обробки документів зі складною версткою застосування виключно базових методів виділення тексту є недостатнім. Використання ієрархічних моделей, зокрема М-Х-У дерев, забезпечує високу точність структурної сегментації та дозволяє зберегти топологічні зв'язки між блоками інформації. Інтеграція подібних алгоритмів у конвеєр систем електронного документообігу суттєво знижує ймовірність структурних помилок при автоматичному розпізнаванні макет.

Список використаних джерел:

1. Кобилін О. А. Методи цифрової обробки зображень : навч. посіб. Харків : ХНУРЕ, 2021. 124 с.