Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

Метод выделения составных частей научного текста на основе анализа распределения пикселей в сканирующей строке

Студент	ИУ7-84Б		К. А. Рунов
	(группа)	(подпись)	(инициалы, фамилия)
Руководитель ВКР			Ю. В. Строганов
		(подпись)	(инициалы, фамилия)
Консультант			
		(подпись)	(инициалы, фамилия)
Консультант			
		(подпись)	(инициалы, фамилия)
Нормоконтролер			А. С. Кострицкий
		(подпись)	(инициалы, фамилия)

2025 год

РЕФЕРАТ

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

1 Аналитический раздел

1.1 Анализ предметной области

1.1.1 Анализ структуры документов (DLA)

Анализ структуры документов (Document layout analysis, DLA) — процесс сегментирования входного изображения документа на однородные компоненты, такие как блоки текста, рисунки, таблицы, графики и т.д., и их классификации [1].

В общем случае анализ структуры документа делится на два взаимосвязанных процесса: физический и логический анализ. Целью физического анализа является выявление структуры документа и определение границ его однородных областей. Целью логического анализа является разметка обнаруженных областей. Выявленные области классифицируются как элементы документа — рисунки, заголовки, абзацы, логотипы, подписи и другие. [2]

Процесс анализа структуры документов состоит из двух основных этапов — этапа предварительной обработки и этапа анализа макета документа [2, 3]. На рисунке ниже приведена схема процесса анализа структуры документов.

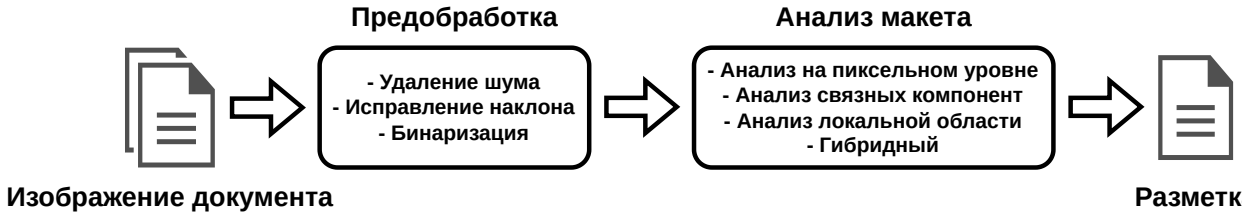

Рисунок 1 – Схема процесса анализа структуры документов [3]

Этап предварительной обработки

Этап анализа макета документа в любом методе анализа структуры документов (далее DLA) часто основывается на определённых предположениях о входных изображениях, таких как отсутствие шума, бинаризация, отсутствие наклона текста или все перечисленные факторы [2, 3].

Цель этапа предварительной обработки — преобразовать входное изоб-

ражение в соответствии с требованиями этапа анализа макета документа конкретного метода [2, 3].

В общем случае на этом этапе используются одна или несколько процедур предварительной обработки, таких как бинаризация, выравнивание и улучшение изображения [2, 4].

Этап анализа макета документа

Анализ макета документа включает в себя определение границ и типов составляющих областей входного изображения документа. Процесс определения границ областей документа называется сегментацией областей документа, а классификация найденных областей по их типу — классификацией областей документа. [3]

Существуют три типа стратегий анализа макета документа: снизу вверх (bottom-up), сверху вниз (top-down) и гибридная (hybrid).

По стратегии снизу вверх (bottom-up) параметры анализа часто вычисляются на основе исходных данных. Анализ макета документа начинается с небольших элементов, таких как пиксели или связанные компоненты. Затем однородные элементы объединяются, создавая более крупные области. Процесс продолжается, пока не будут достигнуты заранее определённые условия остановки.

По стратегии сверху вниз (top-down) анализ макета документа начинается с крупных областей, например, на уровне всего документа. Затем эта большая область разбивается на более мелкие, такие как колонки текста, на основе определённых правил однородности. Анализ сверху вниз прекращается, когда дальнейшее разбиение областей становится невозможным или достигаются условия остановки.

Гибридная стратегия (hybrid) представляет собой комбинацию обеих стратегий (снизу вверх и сверху вниз). [2]

После сегментации областей происходит их классификация с помощью различных алгоритмов, в результате чего формируется логическая структура документа.

По завершении данного этапа извлеченные геометрическая и логическая структуры сохраняются для последующей реконструкции. Для этого, как правило, используется иерархическая древовидная структура данных. [3]

1.1.2 Типы макетов документов

Макеты документов могут иметь различные структуры. Печатные документы можно разделить на шесть типов [5]: прямоугольные, Манхэттенские, не-Манхэттенские, многоколоночные Манхэттенские, с горизонтальным наложением и с диагональным наложением.

Рисунок 2 – Макеты документов: (a) Стандартный (прямоугольный), (b) Манхэттенский, (c) Не-Манхэттенский, (d) Многоколоночный Манхэттенский, (e) Произвольный (сложный), (f) С горизонтальным и диагональным наложением. [2]

На рисунке выше показаны примеры описанных типов макетов документов:

— Стандартный макет характеризуется большими прямоугольными текстовыми блоками, расположенными в одной или нескольких колонках, при этом каждая колонка содержит по одному абзацу.

— Если документ содержит несколько абзацев в колонках, его можно отнести к Манхэттенскому макету. Примеры таких документов — научно-технические статьи, журналы и другие.

— Не-Манхэттенские макеты включают зоны непрямоугольной формы.

— Макеты с наложением содержат элементы, такие как текст, который перекрывает другие элементы документа. Наложение может возникать, например, из-за просвечивания (см. Рисунок 2(f)).

— Документы с произвольными (или сложными) макетами могут включать рукописный и/или печатный текст, содержащий различные стили, типы и размеры шрифтов.

Таким образом, документы, содержащие научно-технические тексты, обычно используют Манхэттенский макет.

1.1.3 Структура научно-технического текста

Научно-технический текст обычно [6, 7, 8] следует четко определенному шаблону и имеет следующую структуру:

1) Название;

2) Информация об авторах;

3) Аннотация и ключевые слова;

4) Введение;

5) Основная часть (кроме текста содержащая в том числе таблицы, рисунки, графики, листинги);

6) Заключение;

7) Ссылки на литературу.

Содержимое научного текста часто не ограничивается текстом, а содержит также следующие составные части:

1) таблицы,

2) листинги,

3) схемы алгоритмов,

4) рисунки,

5) графики.

Зная структуру научного текста и его основные части можно перейти к формализации задачи выделения составных частей научного текста.

1.2 Формализация предметной области

Пусть D — документ, представленный в виде набора изображений, содержащих текст, листинги, таблицы, рисунки и прочие структурные элементы.

Документ

$$D = \{P_1, P_2, \ldots, P_n\}$$

состоит из страниц $P_1, P_2, \ldots, P_n$, а каждая страница P_i в свою очередь содержит множество объектов $O_{i,1}, O_{i,2}, \ldots, O_{i,m}$.

Объект $O_{i,j}$ — кортеж $(x_{i,j}, y_{i,j}, w_{i,j}, h_{i,j})$, где $(x_{i,j}, y_{i,j})$ — координаты верхнего левого угла, $w_{i,j}$ — ширина, $h_{i,j}$ — высота объекта.

Требуется построить отображение

$$F : D \to \{(O_{i,j}, C_{i,j})\},$$

где каждому объекту $O_{i,j}$ ставится в соответствие класс

$$C_{i,j} = C_{i,j}(O_{i,j}),$$

область допустимых значений которого определяется исходя из требований к разметке.

Например, в случае задачи выделения составных частей научного текста, $C_{i,j} \subseteq \{$Фон, Текст, Таблица, Листинг, Схема алгоритма, Рисунок, График, Неопределенность$\}$; Объект классифицируется как «Неопределенность» в случае, когда не удалось распределить его ни в один из предыдущих классов.

Поставленную задачу можно решить, разбив на две подзадачи и решив каждую подзадачу соответственно: первая подзадача — нахождение объектов на страницах и выявление их геометрических свойств, вторая подзадача — классификация найденных объектов (определение $C_{i,j}$ для каждого объекта $O_{i,j}$).

Решением первой подзадачи является построение отображений

$$P_i \to \{O_{i,j}\},$$

решением второй подзадачи является построение отображений

$$O_{i,j} \to C_{i,j}.$$

Далее будут рассмотрены существующие методы, позволяющие решить поставленную задачу.

1.3 Описание существующих методов

1.4 Классификация существующих методов

1.5 Формализованная постановка задачи

Вывод

2 Конструкторский раздел

2.1 Требования и ограничения метода

2.2 Описание разрабатываемого метода

2.3 Тестирование и классы эквивалентности

2.4 Структура разрабатываемого программного обеспечения

Вывод

3 Технологический раздел

3.1 Выбор средств реализации

3.2 Реализация программного обеспечения

3.3 Результаты тестирования

3.4 Пользовательский интерфейс

3.5 Руководство пользователя

Вывод

4 Исследовательский раздел

4.1 Описание исследования

4.2 Результаты исследования

Вывод

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Bhowmik et al. Text and non-text separation in offline document images: a survey // International Journal on Document Analysis and Recognition (IJDAR). 2018. Т. 21.

2. Binmakhashen G.M., Mahmoud S.A. Document Layout Analysis: A Comprehensive Survey // ACM Comput. Surv. 2019. Т. 52, № 6.

3. Bhowmik S. Document Layout Analysis. — Springer Singapore, 2023 — 86 с.

4. Kasturi R., O'Gorman L., Govindaraju V. Document image analysis: A primer // Sadhana — Academy Proceedings in Engineering Sciences. 2002. Т. 27. С. 3–22.

5. Kise K. Page Segmentation Techniques in Document Analysis // Doermann D., Tombre K. Handbook of Document Image Processing and Recognition. — Springer London, 2014. С. 135–175.

6. Бутенко Ю.И. Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. Т. 20, № 1. С. 5–13.

7. Романов Д.А. Кратко о структуре экспериментальной научной статьи на английском языке // Вестник Казанского технологического университета. 2014. Т. 17, № 6. С. 325–327.

8. Раицкая Л.К. Структура научной статьи по политологии и международным отношениям в контексте качества научной информации // Полис. Политические исследования. 2019. № 1. С. 167–181.

ПРИЛОЖЕНИЕ А