**УДК 004.93**
**DOI: 10.15587/2312-8372.2019.173913**

# РАЗРАБОТКА ИЕРАРХИЧЕСКОГО ПОДХОДА К СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ ОТСКАНИРОВАННЫХ ДОКУМЕНТОВ

**Ищенко А. В., Жучковский В. Ю.**

## 1. Введение

Быстрое развитие цифровых технологий привело к переводу в электронный вид всех видов материалов, в том числе документов архивов, библиотек, а также предприятий для создания электронных архивов.

Важным этапом обработки изображений отсканированных документов является сегментация, которая заключается в разделении изображения на однородные области, сходные по одному признаку или набору признаков.

В литературе можно выделить 2 основных подхода к сегментации изображений отсканированных документов [1]: нисходящий и восходящий. Методы сегментации, использующие нисходящий подход [2–4], сначала определяют объекты более высокого уровня структуры страницы – текста и графических элементов, а затем – колонок текста, параграфов, строк, символов текста. Эти методы отличаются высоким быстродействием, но невысоким качеством сегментации, так как не всегда возможно обработать непрямоугольные области текста или заголовки, занимающие несколько колонок текста. Методы сегментации, использующие восходящий подход к сегментации [5, 6], начинают обработку с символов текста, которые объединяются в абзацы, параграфы, колонки. Далее полученные объекты классифицируются как текстовые области, пока не будут выделены цельные области текста. Данные методы отличаются высоким качеством сегментации, так как хорошо обрабатывают изображения со сложной формой, но имеют невысокое быстродействие, в связи с тем, что требуют обработки сначала каждого пикселя, а затем областей документа. При оцифровке большого количества печатных документов возрастают требования к оперативности их обработки. Как следствие актуальной задачей, которая решается в данной работе, является повышение оперативности обработки изображений отсканированных документов путем сокращения времени их обработки, используя иерархический подход к сегментации изображений. Поэтому *объектом исследования* является процесс распознавания областей изображений отсканированных документов. А *целью исследования* является использование иерархического подхода к сегментации изображений отсканированных документов для сокращения времени обработки изображений при достаточном качестве сегментации.

## 2. Методика проведения исследований

Основой любого отсканированного документа является его структура, то есть взаимное расположение графического материала и текста. Области,

включающие однородное содержимое, например, только текст, только графику или только фото, образуют структурные области.

Одним из основных этапов обработки изображений отсканированных документов является сегментация изображений. При сегментации на изображениях выделяются области текста, графики и фото. Каждая из этих структурных областей имеет различные свойства, поэтому сложно подобрать систему признаков для выделения текстовых областей, а также областей графики и фото из фона.

Существующие методы сегментации изображений отсканированных документов не удовлетворяют одновременно требованиям к малому времени обработки при достаточном качестве сегментации.

Необходимость иерархической организации системы необходима в том случае, если для ее реализации требуется затрата большого количества времени, которая недопустима для данной системы. Поэтому для сокращения времени обработки изображения документа в данной работе предлагается использовать иерархический подход к сегментации изображений отсканированных документов. Данный подход заключается в представлении изображения в виде многоуровневой структуры, в которой существует разделение множества составляющих ее объектов на подмножества разных уровней, обладающие свойством целостности [7]. Иерархическая структура позволяет выполнять обработку отдельных информационных массивов. То есть данный подход позволяет выполнять обработку для каждого уровня представления изображения. Согласно иерархическому подходу изображение сначала раскладывается на области иллюстраций, включающие как фото, так и графику, с последующей их классификацией, а также области, содержащие текст и фон, с последующим выделением текстовых областей из фона. В результате сегментации изображения, используя предложенный подход, изображение отсканированного документа представляется в виде отдельных областей: текста, графики, фото и фона. Схема иерархического подхода представлена на рис. 1.

Согласно схеме иерархического подхода к сегментации изображений (рис. 1), первому уровню соответствует исходное изображение отсканированного документа. В качестве базовой модели представления данного изображения выбрана наиболее часто используемая в литературе модель смешанного растрового контента – Mixed Raster Content [8]. Согласно данной модели изображение документа представляется в виде изображения маски, а также изображений переднего плана и фона. Каждое из этих изображений содержит объекты определенного класса и независимо сжимается определенными кодерами.

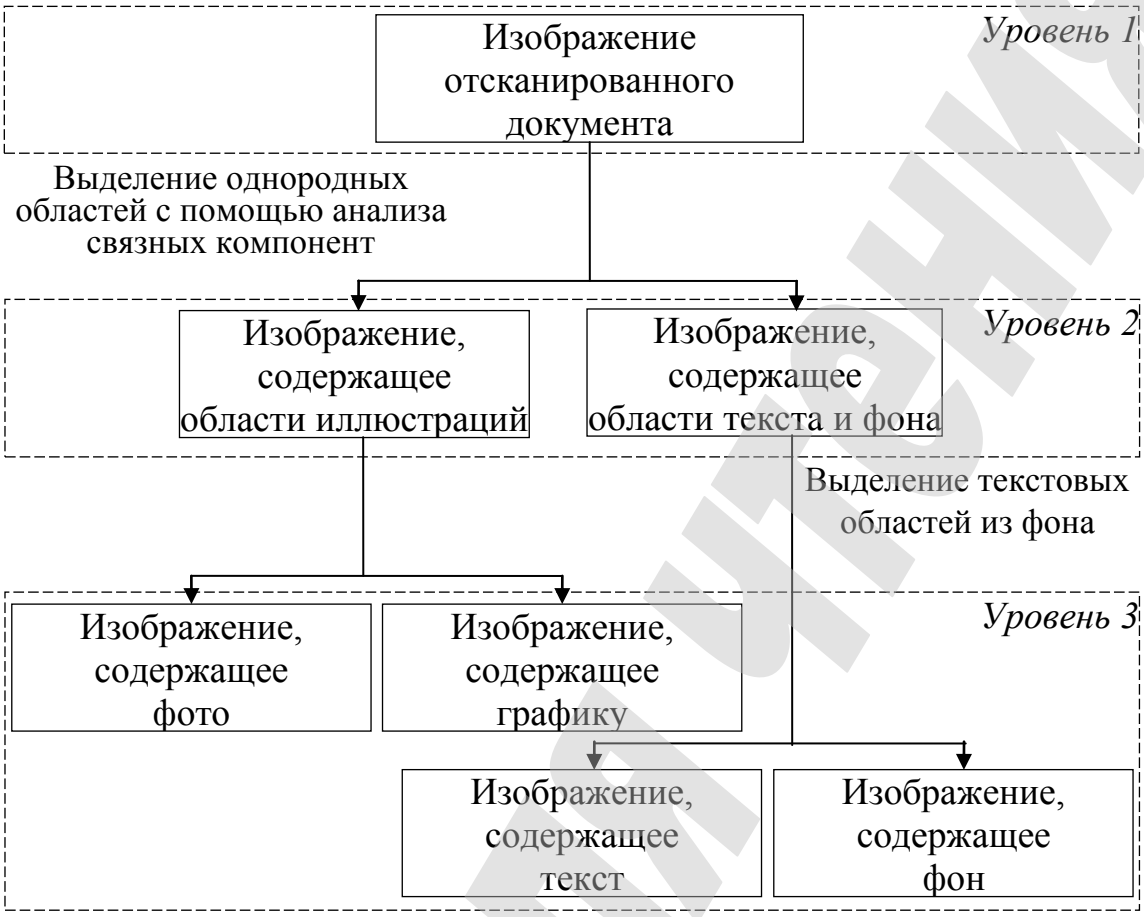


**Рис. 1.** Схема иерархического подхода к сегментации изображений отсканированных документов

Маска содержит информацию о взаимном расположении объектов переднего плана и фона на изображении. Для того, чтобы иметь возможность извлекать информацию из структурных областей, необходимо представление изображения, которое представляет эти структурные области на отдельных изображениях. Поэтому в данном исследовании в качестве представления изображения используется модель работы [9]. Данная модель отличается тем, что она представляет изображение отсканированного документа в виде набора изображений, каждое из которых содержит один класс однородной области – графику, или фото, или текст на однородном фоне. Данная модель изображения представляет области текста в виде структурной текстуры с символами текста в качестве непроизводных элементов, а области иллюстраций – в виде областей постоянной интенсивности. Такое представление изображения отсканированного документа позволяет учесть структурные свойства однородных областей и выбрать систему признаков для их идентификации, которые имеют высокую вычислительную эффективность, что влечет за собой сокращение времени обработки изображения при достаточном качестве сегментации.

Следующим шагом является разделение изображения отсканированного документа на изображение, содержащее иллюстрации, и изображение, содержащее текстовые области и фон. Для этого используется разработанный в [10] метод выделения иллюстраций на изображении отсканированного

документа с использованием усредняющей фильтрации. Согласно этому методу однородные области текста и иллюстраций выделяются с помощью анализа связных компонент. Использование анализа связных компонент позволяет отличить области иллюстраций от областей текста, так как они связные компоненты, которые соответствуют символам текста, отличаются от связных компонент, соответствующих иллюстрациям, своей формой и периодичностью. То есть в качестве признака для идентификации текста выступает размер связных компонент, соответствующих символам текста. Использование иерархического подхода к сегментации позволяет вычислить простые, в вычислительном смысле, признаки для текстовых областей и сегментировать исходное изображение отсканированного документа на 2 изображения: одно из них содержит иллюстрации на однородном фоне, другое – текст и фон. Использование «простых» признаков сокращает время обработки изображения, а использование усредняющей фильтрации при выделении областей иллюстраций сохраняет достаточно высокое качество сегментации изображения.

Согласно иерархическому подходу (рис. 1) на третьем уровне иерархии представлены изображения, содержащие фото, графику, текст, фон, отдельно друг от друга.

Сначала области иллюстраций разделяются на однородные области, которые отличаются своей структурой: фото и графику. Для этого рассматривается часть иерархической структуры, в которой находится изображение, содержащее области иллюстраций. Для этого в [11] разработан метод, предлагаемый для идентификации областей графики и фото с использованием статистическо-геометрических признаков. Для сокращения времени обработки изображения при идентификации областей фото и графики используется разбиение областей иллюстраций на блоки фиксированного размера. При блочной обработке обычно качество сегментации снижается, поэтому необходимо использовать такие признаки идентификации однородных областей, которые обладают высокой вычислительной эффективностью, то есть обеспечивают достаточное качество сегментации при малом времени обработки изображения. Поэтому в качестве признаков идентификации в [11] предложено использовать соотношение размеров объектов и оценку математического ожидания высоты перепада интенсивности на границах однородных областей. В качестве первого выбран признак, который характеризует объект, представляющий графику, как изображение, которое содержит линейные объекты [12], представленные контурами разной ширины и протяженности. Второй признак позволяет отличить области графики от фото, учитывая, что области графики характеризуются меньшим количеством градаций серого, чем фото, что влечет за собой увеличение высоты перепадов интенсивности между градациями.

Далее текстовые области на изображениях отсканированных документов отделяются от фона. Для этого используется методика выделения текстовых областей [13], которая позволяет учесть расстояния между текстовыми символами при обработке строк изображения и расстояния между строками текста при обработке изображения по столбцам. Обработка в окрестности каждого пикселя изображения позволяет сократить время обработки и

сохранить достаточно высокое качество изображения.

### 3. Результаты исследований и обсуждение

Для оценки времени обработки изображений отсканированных документов, используя иерархический подход к сегментации изображений, использовались изображения отсканированных статей и журналов базы данных документов MediaTeam Oulu [14].

Экспериментальные исследования проводились с использованием компьютера, который имеет следующие технические и системные характеристики: процессор Intel Core i5-3210, 2.5 GHz CPU, оперативная память 6GB, 64-разрядная операционная система Windows 7. Было отобрано 92 тестовых изображения отсканированных документов, которые содержали основной текст, заголовки, фото и графику. Изображения используемой базы документов имеют размер 3200×2300 пикселей и отсканированы с разрешением 300 dpi, что дает довольно высокое качество отсканированного документа. Примеры анализируемых изображений представлены на рис. 2.

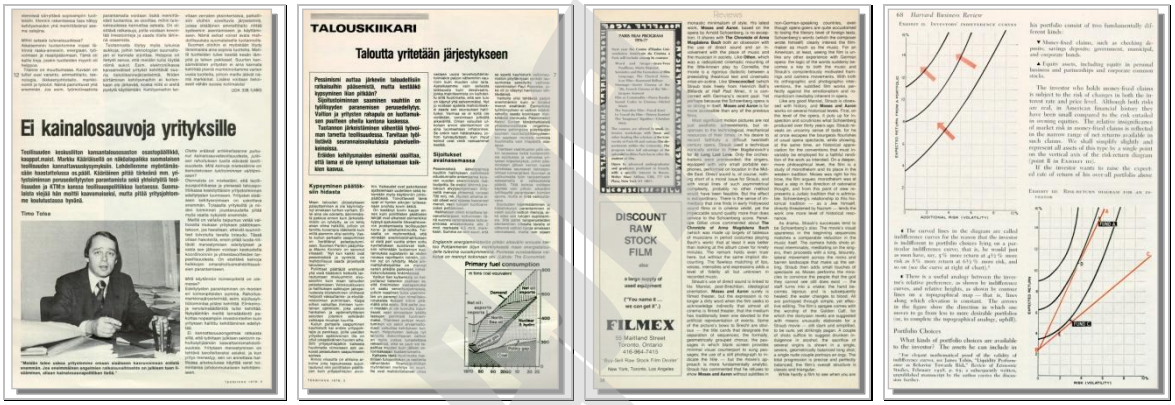


**Рис. 2.** Примеры изображений отсканированных документов базы данных MediaTeam Oulu [14]

Результаты эксперимента сравнивались с результатами работ [2, 3, 5], в которых исследования производились, используя двухъядерный процессор 2.4 GHz CPU и 64-разрядная операционная система Windows 7. Технические и системные характеристики данного исследования и анализируемых работ сопоставимы, поэтому результаты этих работ могут быть добавлены в сравнение.

Результаты по времени обработки для использования иерархического подхода к сегментации и известных подходов приведены в табл. 1.

**Таблица 1**

Время обработки изображений анализируемыми методами

| Метод cluster center-based classification [2] нисходящий подход | Метод Fisher classifier [2] нисходящий подход | Метод работы [3] нисходящий подход | Метод работы [5] восходящий подход | Иерархический подход |
|---|---|---|---|---|
| 158 с | 139 с | 286 с | 155 с | 2,3 с |

Системные и аппаратные ресурсы компьютеров, которые использовались для экспериментов в работах [2, 3, 5] и в данном исследовании имели сопоставимые характеристики. Результаты, представленные в табл. 1, позволяют сделать вывод, что использование предложенного иерархического подхода к сегментации изображений отсканированных документов имеет значительные преимущества в быстродействии.

Дальнейшие исследования могут быть направлены на решение задачи сегментации с целью увеличения быстродействия путем введения параллельной обработки однородных областей на одном уровне структуры.

**4. Выводы**

В работе предложено при сегментации изображений отсканированных документов использовать иерархический подход. Данный подход позволяет однородные области изображения представить с помощью древовидной структуры, в которой области нижнего уровня строго соотносятся с определенной областью изображения верхнего уровня. Согласно этому подходу, области иллюстраций и области, содержащие текст и фон составляют объекты одного уровня структуры. А области фото, графики, текста и фона представляются на другом уровне. Это позволяет выполнять обработку отдельно для каждой области изображения, а также существенно сократить время обработки на каждом уровне.

**Благодарности**

Автор выражает благодарность и глубокую признательность за ценные и конструктивные советы и комментарии при работе над данным исследованием коллегам с кафедры прикладной математики и информационных технологий Одесского национального политехнического университета (Украина):

– доктору технических наук, доценту Поляковой М. В.;

– доктору технических наук, профессору Крылову В. Н.

**Литература**

1. Shafait, F., Keysers, D., Breuel, T. M. (2008). Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (6),* 941–954. doi: http://doi.org/10.1109/tpami.2007.70837

2. Kumar, S., Gupta, R., Khanna, N., Chaudhury, S., Joshi, S. D. (2007). Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model. *IEEE Transactions on Image Processing, 16 (8),* 2117–2128. doi: http://doi.org/10.1109/tip.2007.900098

3. Acharyya, M., Kundu, M. K. (2001). Multiscale Segmentation of Document Images Using M-Band Wavelets. *Lecture Notes in Computer Science,* 510–517. doi: http://doi.org/10.1007/3-540-44692-3_62

4. Cesarini, F., Gori, M., Marinai, S., Soda, G. (1999). Structured document segmentation and representation by the modified X-Y tree. *Proceedings of the Fifth*

*International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No.PR00318)*, 563. doi: http://doi.org/10.1109/icdar.1999.791850

5. Baird, H. S., Moll, M. A., An, C., Casey, M. R. (2007). Document image content inventories. *Document Recognition and Retrieval XIV*. doi: http://doi.org/10.1117/12.705094

6. Вилькин, А., Егорова, М. (2010). Сегментация отсканированных документов. *Материалы конференции «ГрафиКон'2010»*, 339–341.

7. Моисеев, Н. Н. (1981). *Математические задачи системного анализа*. Москва: Наука, 487.

8. de Queiroz, R. L., Buckley, R. R., Xu, M. (1999). Mixed Raster Content (MRC) Model for Compound Image Compression. *Visual Communications and Image Processing*. San Jose, 3653, 1106–1117. doi: http://doi.org/10.1117/12.334618

9. Ishchenko, A., Polyakova, M., Kuvaieva, V., Nesteryuk, A. (2018). Elaboration of structural representation of regions of scanned document images for MRC model. *Eastern-European Journal of Enterprise Technologies, 6 (2 (96))*, 32–38. doi: http://doi.org/10.15587/1729-4061.2018.147671

10. Polyakova, M., Ishchenko, A., Huliaieva, N. (2018). Document image segmentation using averaging filtering and mathematical morphology. *Telecommunications and Computer Engineering (TCSET)*. Lviv-Slavske. doi: http://doi.org/10.1109/tcset.2018.8336354

11. Polyakova, M., Ishchenko, A., Volkova, N., Pavlov, O. (2018). Combined method for scanned documents images segmentation using sequential extraction of regions. *Eastern-European Journal of Enterprise Technologies, 5 (2 (95))*, 6–15. doi: http://doi.org/10.15587/1729-4061.2018.142735

12. Magnier, B., Montesinos, P., Diep, D. (2011). Ridges and Valleys Detection in Images Using Difference of Rotating Half Smoothing Filters. *Lecture Notes in Computer Science*. Ghent, 261–272. doi: http://doi.org/10.1007/978-3-642-23687-7_24

*13.* Гусак, Д. Е., Ищенко, А. В. (2019). Выделение текстовых фрагментов на изображении отсканированного документа. *IX Международная научная конференция «Сучасні інформаційні технології», г. Одесса, Украина, 23–24 мая 2019 г.*, 1–2.

14. Sauvola, J., Kauniskangas, H. (1999). *MediaTeam Document Database II, a CD-ROM collection of document images*. University of Oulu. Finland.