На правах рукописи

ЗАВАЛИШИН Сергей Станиславович

АЛГОРИТМЫ АНАЛИЗА И ОБРАБОТКИ ИЗОБРАЖЕНИЙ СКАНИРОВАННЫХ ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Специальность 05.13.01 – «Системный анализ, управление и обработка информации (технические системы)»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Рязань 2016

Работа выполнена на кафедре автоматики и информационных технологий в управлении ФГБОУ ВО «Рязанский государственный радиотехнический университет»

Научный руководитель: **Бехтин Юрий Станиславович**
доктор технических наук, профессор,
профессор кафедры автоматики и информационных технологий в управлении ФГБОУ ВО «Рязанский государственный радиотехнический университет»

Официальные оппоненты: **Приоров Андрей Леонидович**
доктор технических наук, профессор,
доцент кафедры динамики электронных систем ФГБОУ ВО «Ярославский государственный университет им. П.Г. Демидова»

Кружилов Иван Сергеевич
кандидат технических наук,
инженер первой категории
АО «НПП «Геофизика-космос»

Ведущая организация: **ФГБОУ ВО «Вятский государственный университет»**

Защита состоится 21 декабря 2016 г. в 11 ч. 00 мин. на заседании диссертационного совета Д 212.211.01 при ФГБОУ ВО «Рязанский государственный радиотехнический университет» по адресу: 390005, г. Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Рязанский государственный радиотехнический университет» и на сайте www.rsreu.ru.

Автореферат разослан « » _____ 2016 г.

Ученый секретарь
диссертационного совета Д 212.211.01

кандидат технических наук, доцент  Пржегорлинский В.Н.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Интенсивное развитие информационных систем в последние десятилетия позволило в значительной степени автоматизировать процессы создания и обработки документов, необходимых для функционирования государственных органов и коммерческих предприятий. Обработка изображений сканированных документов представляет собой комплексный процесс, включающий в себя не только изменение самого изображения, но и извлечение большого количества информации о цвете, типе документа, содержании и графике. Интернет-сервисы, например Evernote, WizNote, Google Drive и их аналоги, позволяют извлекать текст из любой фотографии, содержащей буквенные символы, поэтому крупные и средние компании для поддержки своего документооборота требуют аналогичной функциональности и от информационных систем, что обуславливает встраивание отдельных функций интеллектуальной обработки изображений.

Производители программного обеспечения для информационных систем предоставляют решения подобного рода, однако предлагаемые ими алгоритмы требуют детальной настройки под нужды конкретного заказчика. В научно-исследовательских работах преобладает тенденция к минимизации труда человека путём повсеместного внедрения методов машинного обучения. Исследователи не ставят своей целью создать всеобъемлющую и универсальную технологию обработки изображений сканированных документов, а решают узкие конкретные задачи. Сопряжение множества разнородных алгоритмов в рамках единой технологии приводит либо к падению общей скорости работы за счёт высокой вычислительной сложности, либо к снижению качества обработки изображений из-за слабой согласованности отдельных алгоритмов.

В связи с этим актуальной является проблема разработки и согласования (гармонизации) алгоритмов анализа и обработки изображений сканированных документов. В диссертационной работе предлагаются новые алгоритмы контрастирования, обратного растрирования, маркировки связных компонент, извлечения текста и выделения структурных признаков, дополняющие друг друга с учётом возможностей современного аппаратного обеспечения и особенностей решаемой задачи. Разработанная методика классификации структурных элементов сканированных документов объединяет предложенные алгоритмы, позволяя достичь высоких показателей точности классификации и скорости обработки.

Степень разработанности темы. Решению различных проблем, связанных с обработкой и анализом изображений сканированных документов, посвящено большое количество работ, выполненных отечественными и зарубежными учеными.

Проблема низкого контраста решается путём применения методов адаптивного контрастирования, предложенных И. Сафоновым, Р. Соболем, Л. Тао, Н. Морони, К. Зуйдервельдом и другими. В существующих подходах применяется разделение изображения на локальные области по жестко заданной сетке, что не позволяет корректно изменять контраст документов, состоящих из однородного фона и рисунков.

Другая проблема связана с эффектом муара из-за наложения растровых сеток сканера и принтера. Методам устранения данного эффекта посвящены работы Г. Галло, К. Дабова, И. Курилина, однако в них текстура изображений восстанавливается не полностью.

Методы выделения ключевых признаков текста, рисунков и таблиц, характеризующих документ внутри информационной системы, описаны в статьях А. Гордо, Ф. Цезарини, С. Усилина, С. Чена, Дж. Кумара, Х. Гао, С. Бухари и других. Указанные подходы не способны извлекать текст из сканированных документов низкого разрешения.

Методы классификации документов, основанные на выделенных гетерогенных признаках, предложены М. Дилигенти, Е. Аппиани, Дж. Лиангом, Х. Сако и С. Тейлором. Данные методы разработаны в предположении, что документ имеет однородный фон, следовательно, точность классификации существенно снижается при сложном изображении в качестве фона.

Вышеуказанные алгоритмы созданы для решения узкого класса задач, что приводит к их несогласованности при совместной работе и снижению быстродействия информационной системы. Решением данной проблемы является создание методики согласования алгоритмов, что требует переработки существующих и создания новых алгоритмов, отвечающих одинаковым требованиям.

Цель и задачи работы. Целью диссертации является разработка научно-обоснованного алгоритмического обеспечения информационных систем для анализа и обработки изображений сканированных документов, а также соответствующей методики согласования разработанных алгоритмов, обеспечивающих высокую точность классификации структурных элементов документов при повышенном быстродействии.

Для достижения поставленной цели **решаются следующие задачи**:

1. разработка теоретических положений и методики гармонизации алгоритмов обработки изображений сканированных документов;
2. разработка алгоритма адаптивного локального улучшения контраста изображений сканированных документов;
3. разработка алгоритма обратного растрирования печатных изображений;
4. разработка алгоритмов поиска текстовых областей, таблиц, маркировки связных компонент и извлечения признаков изображений сканированных документов;

5. разработка на основе предложенной методики гармонизации алгоритма классификации структурных элементов изображений документов, объединяющего полученные алгоритмы в виде единой технологии анализа и обработки изображений сканированных документов.

Научная новизна. В диссертационной работе получены следующие новые научные результаты.

1. Предложена методика гармонизации алгоритмов обработки изображений сканированных документов на основе многократного повторного применения результатов работы алгоритмов и полутоновых изображений.
2. Разработаны алгоритм адаптивного локального улучшения контраста изображений сканированных документов, использующий сглаживание параметров кривых преобразования между соседними областями с помощью графа связности, и метрика сравнения структурной схожести изображений, полученная путем модификации метрики SSIM.
3. Разработан алгоритм восстановления растрированных изображений на основании модели оператора растрирования и предложен оригинальный подход к сравнению восстановленных изображений.
4. Разработаны алгоритмы поиска текстовых областей, таблиц и маркировки связных компонент на изображениях сканированных документов, использующие дескриптор длин полутоновых отрезков, структурный тензор и карту смежности блоков изображения.
5. Разработаны алгоритмы извлечения ключевых признаков структурных элементов сканированных документов с помощью длин полутоновых отрезков, пространственного локального двоичного шаблона, векторов Фишера на основе распределения Бернулли и алгоритм классификации структурных элементов документов, использующий двухступенчатый подход к объединению нескольких классификаторов.

Методы достоверности исследования. Для решения поставленных задач в работе использовались элементы теории вероятностей и математической статистики, теории оптимального оценивания и фильтрации, численные методы вычислений. Оценка качества работы алгоритмов проводилась на основе статистического моделирования на ЭВМ путем сравнения с существующими аналогами на репрезентативных тестовых выборках.

Практическая значимость. Разработанные алгоритмы обработки и анализа изображений документов и алгоритм классификации, использующий набор гармонизированных алгоритмов, внедрены компанией ООО «Исследовательский центр Самсунг» в качестве программно-аппаратного комплекса, написанного на языке стандарта C++11 с применением библиотеки OpenCV 3.0.0 для загрузки и манипуляции изображениями, что подтверждается соответствующим актом внедрения.

Апробация работы. Основные результаты работы докладывались и получили положительную оценку на международных и российских научно-

технических конференциях: Color Imaging XX: Displaying, Processing, Hardcopy, and Applications, San-Francisco, USA, 2015; XII МНТК “Распознавание-2015”, Курск, 2015; Mediterranean Embedded Computing MECO’2015, Budva, Montenegro, 2015; Visual Information Processing and Communication VII, San-Francisco, USA, 2016; Mediterranean Embedded Computing MECO’2016, Bar, Montenegro, 2016.

Публикации. По теме диссертации опубликовано 8 работ, в том числе 1 патент, 2 работы в журналах ВАК, 3 работы изданы на английском языке и индексированы в базах IEEE и SCOPUS, 1 работа в международном сборнике на английском языке и 1 работа в тезисах конференций. Одна публикация отмечена дипломом за лучшую научную работу.

Личный вклад автора. Все основные результаты диссертации, опубликованные в приведенных работах в составе коллектива авторов – методика гармонизации, алгоритмы контрастирования, обратного растрирования, нахождения текста, таблиц и изображений, извлечения признаков и классификации сканированных документов и сравнение данных алгоритмов с существующими аналогами – получены автором лично.

Объем и структура работы. Диссертационная работа состоит из введения, четырех глав и заключения. Работа содержит 140 страниц основного текста, в том числе 35 рисунков, 13 таблиц, список использованных источников из 233 наименований.

ОСНОВНЫЕ ПОЛОЖЕНИЯ, ВЫНОСИМЫЕ НА ЗАЩИТУ

1. Методика гармонизации алгоритмов обработки и анализа изображений сканированных документов, позволяющая достичь высоких целевых показателей работы информационной системы при относительно малом времени обработки каждого документа.
2. Алгоритм адаптивного локального контрастирования изображений сканированных документов с использованием предварительной сегментации и алгоритм восстановления растрированных изображений на основе методов разреженного кодирования, позволяющий снизить ошибку восстановления на 7 % и повысить число похожих дескрипторов на 5 % в сравнении с известными алгоритмами.
3. Алгоритмы нахождения текста и таблиц на изображениях сканированных документов с помощью структурного тензора и адаптивной бинаризации, достигающие точности 84 % по метрике F1, и алгоритмы извлечения ключевых признаков изображений сканированных документов с использованием пространственного локального двоичного шаблона, гистограммы длин полутоновых отрезков и векторов Фишера на основе распределения Бернулли, позволяющие повысить точность классификации на 10 % в сравнении с известными алгоритмами.

4. Алгоритм классификации структурных признаков изображений сканированных документов с помощью нескольких классификаторов, основанный на разработанной методике гармонизации и повышающий скорость обработки изображений до 2,2 раз в сравнении с известными алгоритмами объединения классификаторов.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** *обоснована актуальность выбранной темы диссертационного исследования, определены основные цели и задачи и приведено краткое изложение содержания работы.*

В **первой главе** *проанализирована типовая схема работы с печатными документами в информационных системах*, состоящая из следующих этапов: 1) захват множества изображений сканированных документов $D = \{D_1 \dots D_n\}$; 2) обработка изображений D с помощью алгоритмов улучшения контраста $f_{enh}: D \to I \in \mathbb{R}$, где $I = \{I_1 .. I_n\}$ – обработанные изображения; 3) поиск областей с однотипным содержимым $Q = \{Q_1 \dots Q_m\} \in I$ и извлечение текста с помощью классификатора $f_{txt}: Q \to \{0,1\}$, где 1 соответствует текстовой области, а 0 – не текстовой; 4) выделение таблиц $T = \{T_1 \dots T_k\} \in I$; 5) Выделение рисунков $I' = \{I'_1 \dots I'_l\} \in I$; 6) извлечение признаков изображений сканированных документов с помощью дескрипторов $d = \{d_1 \dots d_t\}$ таких, что $d: D_j \cap \{I', Q, T\} \to X_j \in \mathbb{R}^N, j = 1..n$, где X_j – признаковое описание изображения документа D_j, и 7) классификация изображений сканированных документов с помощью классификаторов $f_1 \dots f_n: X_j \to \{p_j(i) | i = 1..C\}$, объединённых в решающую систему $F: f_1(X) \dots f_n(X_j) \to C_j \in \mathbb{Z}$, где $p_j(i)$ – вероятность принадлежности изображения документа D_j классу i, C – общее число классов, а C_j – спрогнозированный класс изображения документа D_j.

Проведён анализ алгоритмов обработки изображений сканированных документов, использующихся на каждом этапе: локального улучшения контраста, обратного растрирования, поиска текста и таблиц, маркировки связных компонент, выделения ключевых признаков и классификации изображений сканированных документов. Показана необходимость разработки методики гармонизации вышеуказанных алгоритмов и разработки новых гармонизированных алгоритмов для достижения высокой точности классификации изображений сканированных документов, скорости работы информационной системы и малых искажений изображения при контрастировании и обратном растрировании.

Во **второй главе** *разработана схема работы гармонизированных алгоритмов обработки изображений документов в информационной системе,* опирающаяся на предложенную методику гармонизации.

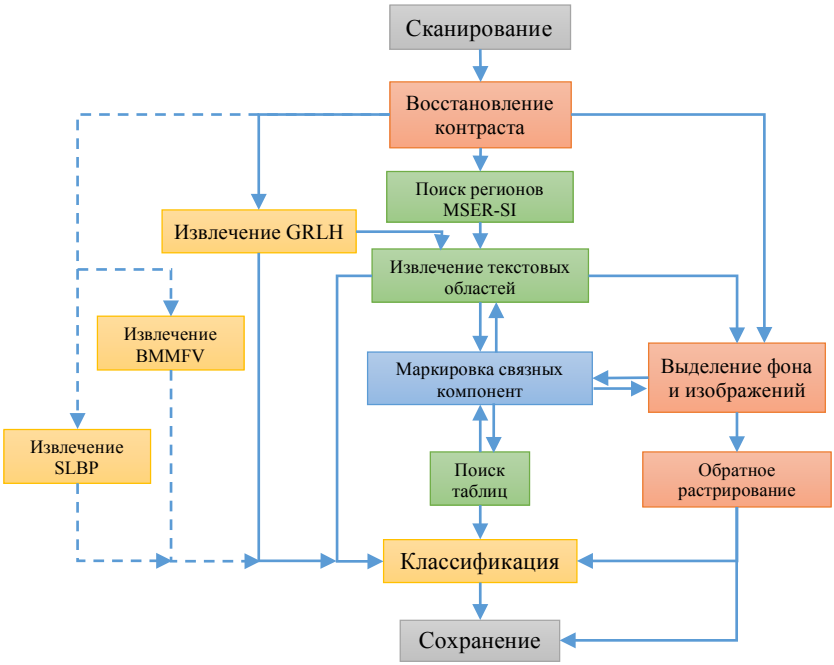

Рисунок 1. Схема работы гармонизированных алгоритмов обработки изображений документов в информационной системе.

Сформулированы следующие основные положения методики.

1. Отсутствие предварительной бинаризации.
2. Многократное применение результатов работы алгоритмов в различных частях информационной системы.
3. Использование высокоскоростных алгоритмов, совокупность которых снижает время работы системы при сохранении точности классификации.
4. Формальная независимость алгоритмов друг от друга, позволяющая удалить или заменить алгоритм без ущерба для работы системы в целом.
5. Использование минимально необходимого набора алгоритмов для обработки конкретного изображения документа.

Разработан алгоритм локального улучшения контраста путём применения локальных S-образных кривых к смежным однородным областям изображения. Схожесть между соседними областями оценивается на основании метрики:

$$S(r_i, r_j) = \alpha_c S_c(r_i, r_j) + \alpha_s S_s(r_i, r_j) + \alpha_f S_f(r_i, r_j), \quad (1)$$

где $S_c(r_i, r_j) = \sum_{k=1}^{n} \min(c_i^k, c_j^k)$ – мера схожести гистограмм областей, $S_s(r_i, r_j) = 1 - \frac{A(r_i)+A(r_j)}{A(I)}$ – мера схожести областей по размеру, а $S_f(r_i, r_j) =$

$1 - \frac{BB(r_i \cup r_j) - A(r_i) - A(r_j)}{A(I)}$ – мера вложенности регионов. Здесь r_i и r_j – i-я и j-я смежные области изображения, c_i^k и c_j^k – k-е ячейки гистограмм областей, $A(\cdot)$ – площадь области, $BB(\cdot)$ – ограничивающий прямоугольник области, а $\alpha_c, \alpha_s, \alpha_f$ – веса каждой компоненты. Гистограммы областей преобразуются с помощью кубического сплайна Эрмита.

Начальная и конечная точки сплайна Эрмита P_{0x} и P_{lx} определяются путём вычисления глобального контраста изображения:

$$P_{0x} = \min\left(D, \min\{i | H[i] \geq H_0\}, \min\left\{i \middle| \sum_{k=0}^{i} H[k] \geq C_0\right\}\right), \quad P_{0x} = \max\left(\max\{\mathrm{i} | \mathrm{H}[i] \geq \mathrm{H}_1\}, \max\left\{\mathrm{i} \middle| \sum_{k=i}^{255} H[k] \geq C_1\right\}\right), \tag{3}$$

где H_0 и H_1, C_0 и C_1 и D – пороги, устанавливающие максимальный уровень гистограммы, площадь гистограммы и интенсивность гистограммы, соответственно. Коэффициенты угла наклона Q_{0x} и Q_{lx} вычисляются с использованием порога бинаризации K, найденного алгоритмом Отцу, и параметра алгоритма $A \in [1;\ 6]$:

$$\text{если } (K_i > 0.5)\ Q_{0x} = 1 + A \times (K_i - 0.5); Q_{0y} = 0; Q_{1x} = 1; Q_{1y} = 0; \quad \text{иначе } Q_{0x} = 1; Q_{0y} = 0; Q_{1x} = 1 + A \times (0.5 - K); Q_{1y} = 0. \tag{4}$$

Полученные коэффициенты K формируют взвешенный граф $G = (V := \{P_{0x}, P_{1x}, K\}, E := S_A)$, вершинами которого являются значения найденных коэффициентов, а рёбрами – коэффициенты схожести соседних областей. Значения коэффициентов сглаживаются между похожими областями для предотвращения резких переходов, где N – количество областей:

$$P_{0x}(r_i) = \frac{1}{\mathrm{N}} \sum_{j=1}^{N} S(r_i, r_j) \cdot P_{0x}(r_j), \quad P_{1x}(r_i) = \frac{1}{\mathrm{N}} \sum_{j=1}^{N} S(r_i, r_j) \cdot P_{1x}(r_j), \quad K(r_i) = \frac{1}{\mathrm{N}} \sum_{j=1}^{N} S(r_i, r_j) \cdot K(r_j). \tag{5}$$

Коэффициенты интерполируются на пиксельную сетку изображения таким образом, что каждому пикселю оказывается сопоставлено собственное значение K, задающее индивидуальное преобразование вида:

$$O = LUT\left(\left[\frac{I - P_{0x}}{(P_{1x} - P_{0x})}\right], K\right), \tag{6}$$

где I – входное значение яркости, $[\cdot]$ – оператор округления, а $LUT(\cdot, K)$ – функция, возвращающая выходное значение яркости из заранее вычисленной таблицы преобразования яркостей.

Разработана метрика оценки структурной схожести цветных изображений:

$$CSSM = \frac{2 \cdot \Delta I \cdot \Delta S}{\Delta I + \Delta S}, \tag{7}$$

где $\Delta S = \frac{\sigma_{L(x)L(y)} + \mathrm{C}}{\sigma_{L(x)} + \sigma_{L(y)} + \mathrm{C}}$, $\Delta I = 1 - [\|LST^*(\bar{x}) - LST^*(\bar{y})\|]^{\alpha}$, $LST^*(\cdot)$ – преобразование цветного изображения в предложенное пространство LST*, $L(\cdot)$ – канал L пространства LST и $\alpha = 0{,}45$ задаёт чувствительность метрики.

Показано, что разработанный алгоритм превосходит существующие алгоритмы повышения контраста по среднему контрасту Михельсона (МС), метрике CSSM и количеству артефактов изображения (таблица 1).

Таблица 1. Сравнение алгоритмов локального контрастирования.

Алгоритм	Средний МС	Мин. МС	SSIM	CSSM	Артефакты
CLAHE	0.78	0.14	**0.92**	0.87	Пятна
MSR	0.82	**0.17**	0.89	0.86	Ореолы
Разр. алгоритм	**0.83**	0.16	0.87	**0.88**	**Нет**

Разработан алгоритм обратного растрирования на основе разреженного кодирования. Сформулирована задача обратного растрирования:

$$\text{найти } \arg\min_{\alpha} \left|\left|GHID_L\alpha - Y\right|\right|^2 + \lambda\|\alpha\|_0 \text{ при } \alpha \in \mathbb{R}^N. \qquad (8)$$

Здесь D_L – словарь низкого разрешения, α – веса элементов словаря, λ – коэффициент регуляризации, а GHI – введённые операторы растрирования, опирающиеся на модель $Y = GHI\bar{X} + \sigma_S + \sigma_P$, где Y – растрированный сигнал, $\bar{X}$ – исходный сигнал, обработанный усредняющим фильтром, G – оператор сглаживания, H – оператор растрирования, I – оператор оценки яркости, а σ_S и σ_P – шумы, вносимые сканирующим устройством и устройством печати.

Решение задачи (8) находится с применением словарей высокого и низкого разрешения D_H и D_L с помощью разработанного алгоритма.

0: Обучается исходный словарь высокого разрешения D_H.
1: Находится преобразование *GHI*.
Для каждого блока *Y* растрированного изображения:
2: находятся веса α путём решения (8);
3: восстанавливается полутоновое изображение $X = D_H\alpha$.

Исходная задача переформулирована в виде следующей задачи:

$$\text{найти } \arg\min_{\alpha} \|ID_L\alpha - (GH)^{-1}Y\| + \lambda\|\alpha\|_0 \text{ при } \alpha \in \mathbb{R}^N, \qquad (9)$$

что позволяет перейти к точкам растра и снизить объём вычислений.

Предложена методика сравнения алгоритмов обратного растрирования, использующая безреференсную оценку резкости BRISQUE, метрику качества MSSSIM и оценку пересечения гистограмм дескрипторов BRISK. Проведено сравнение разработанного алгоритма с существующими в соответствии с разработанной методикой. Преимущество предложенного алгоритма по ряду параметров перед известными алгоритмами отражено в таблице 2.

Таблица 2. Сравнение алгоритмов обратного растрирования

DPI	**Метрика**	**Алгоритм**							
		Исх. изобр.	GS	EPF-I	EPF-II	FFT-D	HFD	TBD	**Разр. Алг.**
600	MS-SSIM	0.29	0.48	0.47	0.40	0.41	0.39	0.46	**0.52**
	BRISQUE	**73.5**	60.7	56.2	39.0	44.5	66.8	53.7	66.2
	BRISK	0.77	0.79	0.86	0.90	0.87	0.92	0.91	**0.94**

В **главе 3** *предложены алгоритмы обнаружения текста и таблиц на изображениях документов и алгоритм маркировки связных компонент изображения на графическом ускорителе.*

Разработан алгоритм поиска текстовых областей, производящий поиск в два этапа: 1) поиск областей с однородным содержимым; 2) классификация

найденных областей на «Текст» и «Не текст». Введена новая формулировка экстремальных регионов Maximal Stable Extremal Regions Supremum and Infimum (MSER-SI; супремум и инфимум наибольших стабильных экстремальных регионов), использующихся для поиска областей с однородным содержимым:

$$\mathrm{Q}' = \{\forall p \in Q, \forall q \in \partial Q : \inf(p) < \inf(q) \text{ или } \sup(p) > \sup(q)\}, \quad (10)$$

позволяющая увеличить число находимых алгоритмом MSER областей. Здесь Q – область изображения, ∂Q – её граница, а p и q – их пиксели.

Разработан дескриптор Grayscale Run Length Histogram (GRLH; гистограмма длин полутоновых отрезков), извлекающий признаки полутонового изображения, который вычисляется следующим образом.

1. Для четырёх направлений D_h, D_v, D_{d+}, D_d (вертикального, горизонтального и двух диагональных), порога T и изображения I находятся отрезки вида:
$$\mathrm{R} = \begin{Bmatrix} a \coloneqq \min(R) \\ b \coloneqq \max(R) \\ \forall i \in [a,b] \end{Bmatrix} \begin{Vmatrix} |I(a-1) - I(a)| > T \\ |I(b+1) - I(b)| > T \\ |I(i) - I(i+1)| \le T \end{Vmatrix}. \quad (11)$$
2. Строятся гистограммы длин найденных отрезков для каждого направления: если длина отрезка $l_i < 128$, то номер ячейки гистограммы, соответствующей данному отрезку, находится как $h_i = \log_2 l_i + \left[c_i / c_{max} \cdot q\right] \cdot h_{max}$, иначе как $h_i = \log_2 l_i + \left[c_{max} \cdot q\right] \cdot h_{max}$, где c_i – средняя яркость отрезка, $c_{max} = 255$, $q = 4$, а $h_{max} = 6$.
3. Полученные гистограммы объединяются в единый вектор признаков размером 96 и нормализуются в диапазоне [0; 1].

Алгоритм MSER-SI был применён для поиска кандидатов в текстовые области, а на основе дескриптора GRLH был построен классификатор найденных областей на «Текст» и «Не текст». Совокупность данных алгоритмов позволила получить более высокую точность нахождения текстовых областей в сравнении с существующим дескриптором RLH, что отражено в таблице 3.

Таблица 3. Сравнение алгоритмов нахождения текстовых областей, %

Алг.	Точн.	Отз.	F1
MSER	82,7	91,7	85,0
MSER-SI	**96,3**	**94,0**	**95,1**

Дескр.	Точн.	Отз.	F1
RLH	**81,3**	80,6	80,1
GRLH	80,4	**85,0**	**83,7**

Разработан алгоритм поиска таблиц на изображениях сканированных документов. Таблица представляется в виде следующих элементов: колонки, ряды, ячейки, пересечения линий и внешние углы. Детектирование линий производится путём вычисления матрицы детектора Харриса размером 2×2. Собственные числа этой матрицы, λ_1 и λ_2, задают условие наличия линии. При $|\lambda_1| \gg |\lambda_2|$, в точке (x, y) присутствует граница с чётко выраженным направлением. Мера выраженности данного направления (яркость границы) оценивается как $c_w = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2$.

Найденные пересекающиеся линии L образуют структуры, вписанные в ограничивающий прямоугольник $R = (R_x, R_y, R_w, R_h)$ с угловыми областями $R' = \{R'_1 \dots R'_4\} \in R$. К таблицам относятся связные компоненты, удовлетворяющие следующим условиям:

$$T = \left\{ CC \left| \begin{array}{c} |C_{ri} \in R| < 5, |C_{oi} \in R'| \geq 4, \\ \min\left(\frac{R_w}{R_h}, \frac{R_h}{R_w}\right) < 0.1, R_w > 50, R_h > 50, \\ \frac{A(CC)}{R_w \cdot R_h} < 0.3, \frac{A(Q_{txt} \in R)}{R_w \cdot R_h} > 0.01 \end{array} \right. \right\}, \quad (12)$$

где CC – связная компонента, $C_r = \{C_{r1} \dots C_{rm}\}$ – углы, $\mathrm{A}(\cdot)$ – площадь, а Q_{txt} – текстовые области.

Разработан алгоритм параллельной маркировки связных компонент на графическом ускорителе, использующий в качестве минимальных элементов блоки размером 2×2 пикселя. Алгоритм разделён на 4 фазы: 1) инициализация карты блоков; 2) сканирование; 3) анализ; 4) финальная маркировка. Общий объём необходимой памяти для работы алгоритма составляет 6,25N байт, где N – число пикселей изображения. Данные хранятся в виде карты блоков, содержащей метку блока и информацию о связности. Инициализация карты блоков производится следующим образом:

```
1:  x, y ← Координаты вычислительного устройства в массиве блоков;
2:  w ← Ширина массива блоков;
3:  Pixels ← Массив пикселей изображения;
4:  bLabels ← Массив меток изображения;
5:  bConn ← Массив шаблонов связности блоков;
6:  P := 0x0;
7:  P0 := 0x777;
8:  Если Pixels[2x, 2y] > 0, то P := P ИЛИ' P0; конец если
9:  Если Pixels[2x+1, 2y] > 0, то P := P ИЛИ' ПСВ(P0, 1); конец если
10: Если Pixels[2x, 2y+1] > 0, то P := P ИЛИ' ПСВ(P0, 4); конец если
11: Если Pixels[2x+1,2y+1] > 0, то P := P ИЛИ' ПСВ(P0, 5); конец если
12: Если P > 0, то
13:     bLabels[x + y·w] := x + y·w + 1;
14:     Если ИмеетБит(P, 0x0) И Pixels[2x-1, 2y-1] > 0, то
15:         УстановитьБит(bConn[x+y·w], 0x0);
16:     конец если
17:     Если (ИмеетБит(P, 0x1) И Pixels[2x, 2y-1] > 0) ИЛИ
              (ИмеетБит(P, 0x2) И Pixels[2x+1, 2y-1] > 0), то
18:         УстановитьБит(bConn[x+y·w], 0x1);
19:     конец если
20:     ...
конец если
```

Функция *ПСВ* осуществляет побитовый сдвиг влево на заданное число бит, функции *ИмеетБит* и *УстановитьБит* проверяют и устанавливают заданные биты, соответственно, а *ИЛИ'* обозначает операцию побитового «ИЛИ».

Оценена сложность алгоритма (количество операций доступа к памяти), которая составила $7{,}25N + 0{,}25M$, где M – максимальная длина цепочки связных пикселей. Разработанный алгоритм модифицирован для маркировки связных компонент воксельного изображения. Необходимый объём памяти модифицированного алгоритма составил $6N$ байт, а сложность алгоритма – $14{,}25N + 0{,}125M$. Оценка времени работы разработанного алгоритма показала преимущество над существующими алгоритмами маркировки, что показано на рисунке 2.

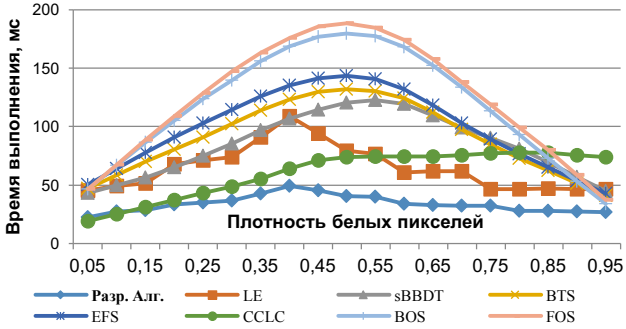

Рисунок 2. Время работы двухмерных алгоритмов маркировки на шумовых структурах с различной плотностью белых пикселей в разрешении 4096×4096, мс.

В **главе 4** *предложен алгоритм классификации изображений документов, опирающийся на методику гармонизации, разработанную во второй главе.* Предложенный алгоритм использует алгоритмы восстановления контраста и обратного растрирования, разработанные во второй главе; извлечения текстовых регионов, таблиц и маркировки связных компонент, разработанные в третьей главе. В соответствии с разработанной методикой гармонизации признаки документа, извлечённые с помощью разработанного в третьей главе дескриптора GRLH, используются повторно для работы классификатора изображений сканированных документов.

Разработан дескриптор Spatial Local Binary Pattern (SLBP; пространственный локальный двоичный шаблон), использующий пространственную пирамиду гистограмм локальных двоичных шаблонов. Разработанный дескриптор масштабирует изображения пространственной пирамиды к размеру 100×100 пикселей, что позволяет повысить точность классификации.

Разработан дескриптор Bernoulli Mixture Model Fisher Vectors (BMMFV; вектора Фишера на основе смеси распределений Бернулли), осуществляющий извлечение векторов Фишера, построенных на основе распределения Бернулли. Использование данного распределения обеспечивает повышение точности классификации при использовании бинарных дескрипторов, например ORB и BRISK.

Разработанные дескрипторы GRLH, SLBP и BMMFV используются для построения трёх независимых классификаторов на основе машин опорных векторов, которые объединяются в единую решающую систему. Разработан алгоритм классификации изображений документов с применением нескольких классификаторов, использующий следующее решающее правило:

$$c(I) = \begin{cases} \arg\max\limits_{i} p_\tau(i|I), & \text{если } p(M_\tau|\Delta P_\tau) > 0,95, \\ \arg\max\limits_{i} p_0(i|I), & \text{если } p(M_0|\Delta P_0) > 0,95, \\ 0, & \text{в противном случае,} \end{cases} \quad (13)$$

где $p_\tau(i|I)$ – вероятность принадлежности изображения I к классу i, оцененная классификатором $C_\tau \in C: \tau = \arg\min\limits_{T} T[C_n(I)]$, $T[\cdot]$ – среднее время классификации изображения документа тестовой выборки, $M_\tau \coloneqq \{i = \arg\max\limits_{i} p_\tau(i|I)\}$, а ΔP_τ – это отношение максимальной вероятности к сумме остальных вероятностей:

$$\Delta P_\tau = \frac{\max\limits_{i} p_\tau(i|I)}{\sum_{i=\{1..K\}\setminus M} p_\tau(i|I)}. \quad (14)$$

Оценка правдоподобия ΔP_τ вычисляется на основе формулы Байеса:

$$p(M_\tau|\Delta P_\tau) = \frac{p(\Delta P_\tau|M_\tau)\, p(M_\tau)}{p(\Delta P_\tau)}. \quad (15)$$

Вероятности принадлежности изображения классам обучающей выборки, оценённые классификаторами $C = \{C_1, \dots, C_N\}$, объединяются в единый вектор и подаются на вход мета-классификатору C_0. Оценка правдоподобия классификации с помощью мета-классификатора $p(M_0|\Delta P_0)$ вычисляется аналогично (15). При условии, что $p(M_0|\Delta P_0) \leq 0,95$, изображение документа помечается как документ неизвестного класса.

Автоматически сгенерирован ряд тестовых выборок, имитирующих реальные документы. Проведено экспериментальное исследование разработанного алгоритма классификации изображений сканированных документов в условиях наличия фона и без него, показавшее его превосходство над существующими алгоритмами (таблица 4).

Таблица 4. Сравнение точности классификаторов изображений документов, %

Выборка	Классификатор				
	LBP	RLH	GMMFV	MV	Разр. Алг.
FlexBack	29,1	81,7	67,2	80,5	**90,8**
FlexBack без фона	48,5	86,4	89,7	90,6	**96,8**

Проведена оценка скорости работы разработанной методики классификации, показавшая её превосходство над существующими методиками объединения классификаторов и некоторыми индивидуальными классификаторами (рисунок 3).

В заключении подведены итоги проделанной работы и сформулированы основные научные и практические результаты.

Прилагаются копии актов о внедрении результатов работы.

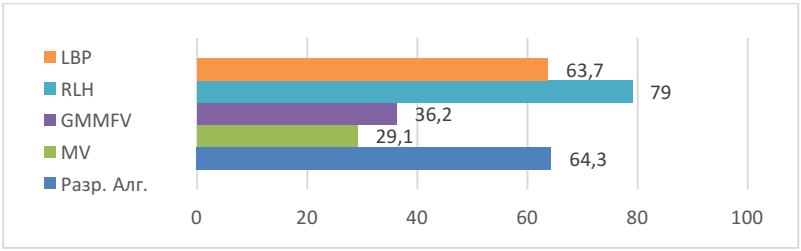

Рисунок 3. Сравнение скорости работы алгоритмов классификации, изобр/с

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Проведён обзор известных методов и алгоритмов анализа и обработки изображений сканированных документов, который показал отсутствие согласованности между применяемыми на практике алгоритмами.
2. На основе обзора предложена методика гармонизации алгоритмов путём многократного повторного применения результатов работы алгоритмов и ориентации на полутоновые изображения. Предложенная методика легла в основу для разработки ряда гармонизированных алгоритмов анализа и обработки изображений сканированных документов, представленных в диссертации.
3. Предложена метрика структурной схожести изображений, позволяющая корректно сравнивать цветные изображения на основании человеческого цветовосприятия.
4. Разработан алгоритм адаптивного локального контрастирования изображений сканированных документов, позволяющий повысить контраст на 1,2 % по метрике Михельсона и снизить искажения на 1,2 % по разработанной метрике структурной схожести.
5. Предложена методика сравнения качества работы алгоритмов обратного растрирования изображений, использующий оценку структурной схожести, безреференсную метрику резкости и схожесть гистограмм дескрипторов изображений.
6. Разработан алгоритм восстановления растрированных изображений на основе методов разряженного кодирования, позволяющий снизить ошибку восстановления на 7 % и повысить число похожих дескрипторов на 5 % при сравнении с помощью предложенной методики.
7. Разработаны алгоритмы поиска текста и таблиц на изображениях сканированных документов с помощью структурного тензора и адаптивной бинаризации, имеющие точность до 84 %, измеренную в соответствии с метрикой F1.
8. Разработан алгоритм маркировки связных компонент с использованием графического ускорителя, позволяющий повысить скорость обработки

изображения до 2,2 раз для двумерного изображения и до 3,2 раз для воксельного изображения.

9. Разработаны алгоритмы извлечения признаков структурных элементов изображений сканированных документов с использованием пространственного локального двоичного шаблона, полутоновых гистограмм длин отрезков и векторов Фишера на основе распределения Бернулли, позволяющие повысить точность классификации до 10 % в сравнении с существующими алгоритмами.
10. Разработан алгоритм классификации структурных элементов изображений сканированных документов с помощью нескольких классификаторов, повышающий скорость обработки изображения до 2,2 раз в сравнении с известными методами при сопоставимой или бóльшей точности.
11. Разработана программная реализация вышеуказанных алгоритмов, позволившая провести статистические исследования на ЭВМ, а также решить вопросы модификации программно-аппаратного комплекса анализа и обработки сканированных изображений компании ООО «Исследовательский центр Самсунг» (г. Москва).

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Работы, опубликованные в научных журналах, входящих в перечень ведущих рецензируемых журналов и изданий ВАК РФ

1. Бехтин Ю.С., Завалишин С.С., Алгоритм эквивалентных отрезков для параллельной маркировки связных компонент бинарного изображения // *Известия ЮЗГУ, сер. "Управление, вычислительная техника, информатика. Медицинское приборостроение",* т. 5, № 56, с. 50-57, 2014.
2. Бехтин Ю.С., Завалишин С.С., Алгоритм параллельной маркировки связных компонент на изображениях, содержащих текст // *Известия ЮЗГУ, сер. "Управление, вычислительная техника, информатика. Медицинское приборостроение",* т. 5, № 50, с. 77-83, 2013.

Работы, опубликованные в сборниках научных трудов на английском языке и индексированные в базах IEEE и Scopus

3. V. Gurov, Y. Bekhtin, S. Zavalishin, A Run Equivalence Algorithm for Parallel Connected Component Labeling on CPU // *Embedded Computing (MECO), 4th Mediterranean Conference*, с. 276-279, 2015.
4. I. Kurilin, I. Safonov, M. Rychagov, S. Zavalishin, D.H. Han, S.H. Kim, Fast algorithm for visibility enhancement of the images with low local contrast // *Proc. SPIE 9395, Color Imaging XX: Displaying, Processing, Hardcopy, and Applications*, 93950B, 2015.
5. S. Zavalishin, Y. Bekhtin, V. Gurov, Inverse Halftonning Using Sparse Coding Methods // *Embedded Computing (MECO), 5th Mediterranean Conference*, с. 381-384, 2016.

Работы, опубликованные в сборниках научных трудов международных конференций на английском языке

6. S. Zavalishin *et al.*, Block Equivalence Algorithm for Labeling 2D and 3D Images on GPU // *Electronic Imaging*, т. 2016, №. 2, с. 1-7, 2016.

Работы, опубликованные в сборниках научных трудов всероссийских конференций на русском языке

7. Завалишин С.С., Бехтин Ю.С., Поиск текста в печатных документах с помощью полутонового дескриптора на основе длин отрезков // *Сборник материалов XII МНТК "Распознавание - 2015"*, Курск, 2015.

Патент на изобретение Российской Федерации

8. Завалишин С.С., Курилин И.В., Система обработки изображений и способ устранения растровой структуры изображения через разреженное представление сканированных печатных копий // *Патент* RU 2595635 C2, 2014.

ЗАВАЛИШИН Сергей Станиславович

Алгоритмы анализа и обработки изображений сканированных документов в информационных системах

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Подписано в печать . Формат бумаги 60x84 1/16.
Бумага ксероксная. Печать ризографическая. Усл. печ. л. 1,0.
Уч.-изд. л. 1,0. Тираж 100 экз. Заказ .
Рязанский государственный радиотехнический университет.
390005, Рязань, ул. Гагарина, 59/1.
Редакционно-издательский центр РГРТУ.