Автоматизированная система реставрации и обработки изображений старопечатных текстов и рукописей

Южиков В.С.
Казанский Государственный Университет

Аннотация

В статье описывается программная система для реставрации изображений старопечатных текстов и рукописей, а также основные подходы и методы для решения данной задачи.

1 Введение

В настоящее время во многих библиотеках и музеях мира хранятся старопечатные книги и рукописи XV-XIX веков. Многие из этих экземпляров бесценны с исторической точки зрения. Но по причине плохого состояния большинства книг и рукописей, доступ к ним, как правило, существенно ограничен для читателей. В то же время, для написания научных работ и проведения исследований необходимо длительное изучение рукописи. Для разрешения данной задачи создаются электронные коллекции старопечатных текстов и рукописей.

Но при оцифровке возникает множество проблем, связанных с плохим качеством многих страниц оригинала. Встречающиеся дефекты можно условно разделить на 2 типа. К первому отнесем дефекты самих страниц книги, появившиеся в результате долгого хранения, действия влажности, температуры, поражения грибком отдельных страниц, выцветание букв, неравномерный цвет бумаги, крупные и мелкие пятна и т.д. Ко второму типу можно отнести дефекты, возникшие при оцифровке, это: неравномерная яркость и контрастность изображения (часто проявляется при съемке цифровым фотоаппаратом), просвечивание надписей с обратной стороны листа, цифровой шум. Все эти дефекты сильно мешают при дальнейшем использовании полученных изображений. В связи с этим большое значение приобретает процесс предварительной обработки изображения.

Многие из дефектов можно устранить при помощи универсальных графических редакторов, типа Adobe Photoshop, но все операции приходится производить вручную, что занимает очень много времени и требует соответствующей квалификации опретора. Поэтому была поставлена задача - создать специализированную систему, в которой были бы реализованы все необходимые функции с возможностью как полностью автоматической работы, так и с поддержкой ручной коррекции процесса обработки.

2 Обзор существующих систем

Сейчас практически отсутствуют программные системы, предназначенные для автоматического и полуавтоматического устранения вышеописанных дефектов. В основном, в публикациях встречаются описания работ, начатых в этом направлении, а также возможные подходы для решения отдельных задач реставрации: [3,5,7].

Описание реализованной системы для комплексного восстановления изображения приведено в [1,2]. Но алгоритмам этой системы присущи некоторые недостатки, а именно выбор конкретной модели изображения (рисунки на рукописи только штриховые, наличие красно-коричневых пятен, более-менее равномерная яркость фона), что не позволяет эффективно обрабатывать широкий класс реальных изображений. Также можно отметить отсутствие весьма важных функций - автоматический поворот изображения, устранение проступания надписей с обратной стороны листа, разметка страницы и т.д.

В связи с этим, разработка и реализация системы реставрации изображений старопечатных текстов и рукописей является весьма актуальной.

3 Обработка оцифрованных изображений

3.1 Постановка задачи

Для дальнейшей автоматической обработки оцифрованных изображений (в частности, для распознавания) старопечатных текстов и рукописей требуется, как правило, преобразовать эти изображения в бинарный вид (это изображение, состоящее только из двух цветов: черного и белого). С другой стороны, электронные коллекции старопечатных текстов создаются и для просмотра их людьми – в этом случае желательно сохранить исходный цвет букв и фона, но при этом очистить изображение от вышеописанных дефектов. Параллельно возникает проблема, связанная с хранением изображений. Желательно, чтобы изображения занимали как можно меньше места при сохранении хорошего качества. Одним из решений является отделение букв, рисунков и других нужных элементов от фона на изображении, который, как правило, не несет большой смысловой нагрузки.

Для решения этих задач и была создана система реставрации и обработки изображений старопечатных текстов и рукописей.

3.2 Нормализация и преобразование изображения в полутоновый черно-белый вид

Первым этапом обработки изображения является нормализация его по яркости и контрастности, т.к. изображения, полученные в результате оцифровки разных книг и разными методами (цифровой фотоаппарат, сканер...), как правило, имеют существенно отличающуюся яркость, контрастность, цветовой баланс. Для размещения таких изображений в рамках одной электронной коллекции желательно, чтобы вышеперечисленные параметры не очень сильно различались.

Для решения этой задачи необходимо перераспределить яркости точек в исходном изображении таким образом, чтобы охватить весь доступный диапазон яркости. Для этого можно использовать разные методы контрастирования. Если результат обработки, после выполнения данной операции, предназначен для просмотра человеком, то целесообразнее применить экспоненциальное контрастирование [6], т.к. оно учитывает нелинейность восприятия глазом различных значений яркости. Если же данная операция, это всего лишь промежуточная обработка для последующих этапов, то лучше использовать метод линейного контрастирования [6].

Если исходное изображение цветное, то, в таком случае, описанному процессу подвергается каждый из трех цветовых каналов изображения. При этом выравнивается

баланс всех цветов, исключая, таким образом, влияние разных условий освещения и особенностей применяемой аппаратуры при съемке. После такой нормализации, изображение переводится в полутоновый черно-белый вид для удобства дальнейшей обработки.

3.3 Склеивание изображения из нескольких фрагментов

Сканирование старопечатных книг и газет большого формата часто приходится делать в несколько приемов, т.к. размеры области сканирования ограничены. В этом случае получается несколько фрагментов изображения, которые потом необходимо склеить в одно большое изображение. Для автоматизации этого процесса был разработан алгоритм, осуществляющий этот процесс. Он основан на поиске похожих элементов в разных фрагментах изображения и дальнейшем объединением по найденным соответствиям элементов. Для успешной работы данного алгоритма необходимо, чтобы разные фрагменты одного изображения были сделаны с «нахлестом», т.е. чтобы на краю каждого фрагмента изображения содержался бы небольшой кусок с другого фрагмента. Пример работы приведен на Рис. 1-2.

Рис. 1. Два исходных изображения

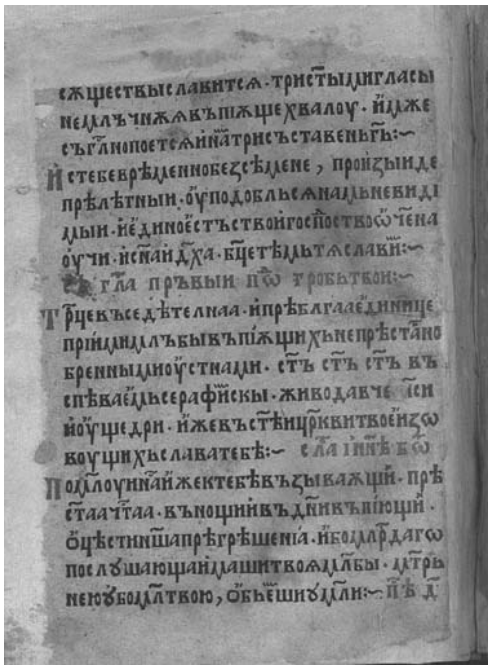

Рис. 2.Результат склейки

3.4 Преобразование в бинарный вид

После вышеописанной предварительной обработки изображения необходимо отделить текст и рисунки от фона. Этот этап является наиболее важным и сложным, т.к., в основном, от него зависит корректность и эффективность полученного при дальнейших преобразованиях результата. После отделения, нужные элементы и фон будут раздельно обрабатываться специальными алгоритмами, что обеспечивает более качественный результат, нежели когда целиком очищается и обрабатывается все изображение. Наиболее удобной формой для представления нужных элементов является бинарный вид – здесь черный цвет будет представлять собой нужные элементы, а белый – фон. Кроме того, бинарный вид часто используется для последующей автоматической обработки изображений, например при распознавании символов. Иногда изображения в таком формате используются и для просмотра их людьми, но такой вариант часто является не совсем удачным, т.к. теряется вся информация о цвете букв, фона и возможна потеря некоторых фрагментов букв. Хотя с другой стороны бинарный вид является наиболее экономичным с точки зрения занимаемого объема.

Методы преобразования в бинарный вид можно условно разделить на 2 типа:

- методы с постоянным порогом преобразования по всему изображению;
- методы с изменяющимся (адаптивным) порогом преобразования.

Первый тип наиболее прост в реализации и быстро работает, но его применение для обработки «плохих» оригиналов дает неудовлетворительные результаты, что иллюстрирует следующий пример:

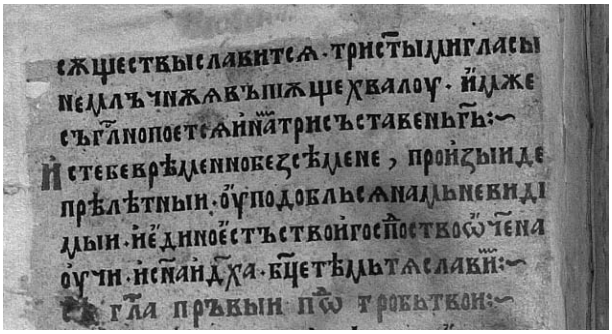

Рис. 3. Исходное изображение (Российская Государственная Библиотека. Часослов. Краков, 1491. Лист 25-оборотный)

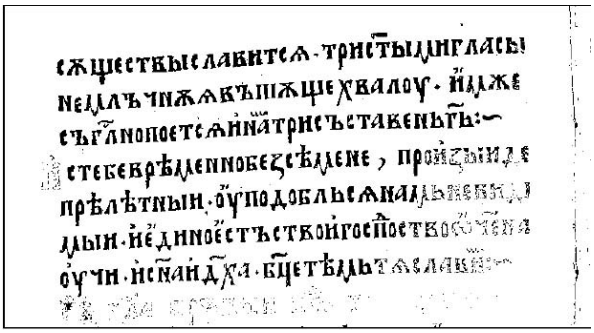

Рис. 4. Низкий порог бинаризации

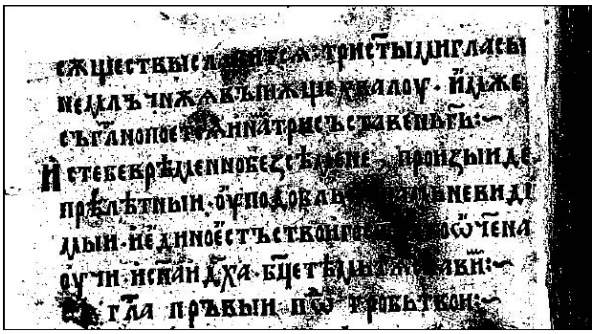

Рис. 5. Высокий порог бинаризации

На Рис. 3-5 хорошо видно, что обеспечить равномерную проработку букв, при наличии пятен и перепадов яркости, нельзя, выбирая постоянный порог по всему изображению. Поэтому целесообразнее применять алгоритм с переменным порогом. Т.к. исходное изображение для бинаризации достаточно специфично, то был разработан алгоритм, учитывающий особенности изображения для получения хороших результатов. Основная идея алгоритма заключается в следующем: изображении разбивается на разные по форме небольшие участки и далее каждый из участков анализируется и для него выбирается оптимальный порог бинаризации - он определяет значение точки в процессе бинарного преобразования, т.е. если яркость данной точки ниже порога, то она считается черной, иначе – белой. В процессе анализа отдельного участка используется предположение, что в пределах этого участка яркость меняется незначительно. При этом учитывается найденные пороги для ближайших участков, чтобы минимизировать возможные ошибки в местах с сильными дефектами изображения. Далее методом бикубической интерполяции находятся и запоминаются пороги преобразования для каждой точки изображения. Результат работы данного алгоритма приведен на Рис. 6.

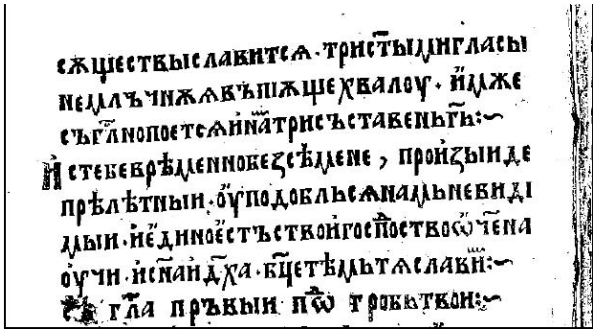

Рис.6. Переменный порог бинаризации

3.5 Поворот изображения

Часто, при сканировании страниц, текст оказывается не параллелен краю сканера. Особенно часто подобное наблюдается при сканировании книг большого формата или при их фотогфировании. Для исправления этого дефекта в систему была введена функция анализа угла наклона текста и поворота изображения, в случае необходимости.

3.6 Фильтрация и очистка от мусора

После выполнения предыдущих этапов было получено бинарное изображение, где черным цветом обозначены нужные элементы (текст, линии рисунков и.т.д.), а белым – фон. Для последующих этапов анализа изображения необходима некоторая информация, а именно

– толщина линий букв на изображении. Для его нахождения был разработан специальный алгоритм – он работает в четыре прохода (по четырем направлениям), на каждом из этих проходов вычисляется толщина линий в текущем направлении, и затем вычисляется результирующее значени тлщиныдля каждой точки изображения. В результате нахождения толщины линий станоится возможным определить оптимальные параметры фильтрации. Если же не использовать информацию о толщине линий на изображении то появляется вероятность удаления элементов букв, либо самих букв в результате фильтрации. Такое может происходить из-за того, что даже одно и то же изображение, отскированое с м зрешением, имеет разные размеры (в пикселях) составляющих его элементов и, если зафиксировать параметры фильтрациионимальные для одного изображения, то при обработке другого изображения (либо этого же изображения, но с другим разрешением) результат будет некорретным.

Для очистки бинарного изображения от шума, мусора и других мелких помех был использован следующий алгоритм: у каждой точки просматривая некотораность (еазмер зависит от толщины линий букв). Если число черных пикселей меньше определенного порога относительно общего количества пикселей в окрестности, то даная точка удаляется (перекрашивается в белый цвет). При там подходе небольшие фрагменты букв не принимаются за "мусор" и не удаляются.

В результате проделанных операций мы отсекли случайные пятна и мусор, ошибочно принятые на нужные элементы. Осталось обработать фон изображения. Для этого используется информация о пороге преобразования для каждой точки, которая была найдена на этапе бинаризации в п.3.4. На Рис.7. показано, как выглядит распределение порога преобразования для изображения на Рис.3.

Рис. 8. Распределение порога преобразования.

Для выравнивания яркости фона и удаления крупных пятен, из значения яркости каждой точки исходного изображения вычитается по модулю соответствующее значение порога преобразования. Далее фон обрабатывается медианным фильтром [4], т.к. он в лучшей степени позволяет сохранить значимые элементы фона.

На заключительном этапе объединяются изображения нужных элементов и изображение фона. На Рис.10. показан конечный результат такой обработки, предназначенный для просмотра человеком.

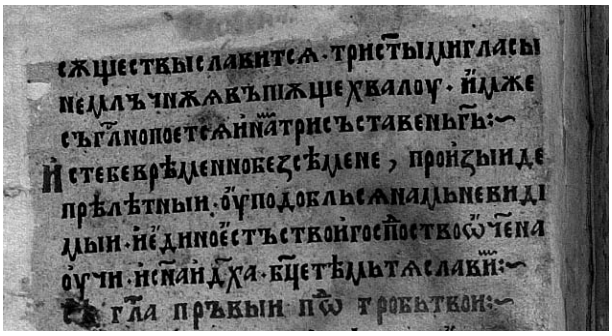

Рис.9. Исходное изображение.

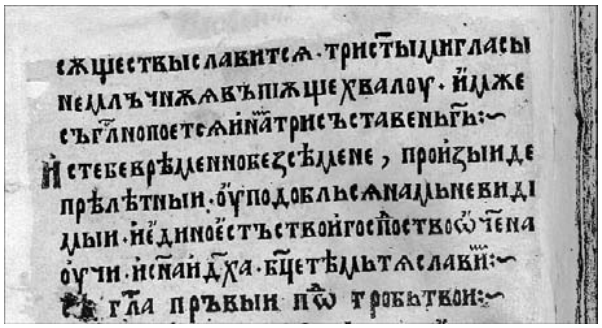

Рис.10. Обработанное изображение.

3.7 Выделение текстовых блоков

Для последующей автоматической обработки и классификации изображения желательно выделить на изображении блоки с текстом. Для выделения используется предположение, что на участках с текстом содержатся более-менее однородные по толщине элементы – отрезки, кривые, точки. При этом применяется информация о найденной толщине линий на изображении. Результат выполнения этапа приведен на Рис.7.

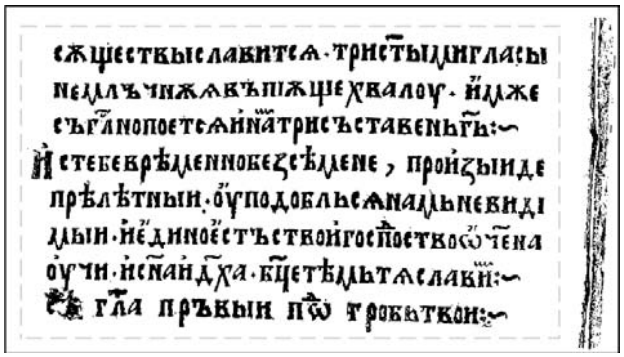

Рис. 7. Выделение текстовых блоков (показано пунктиром)

4 Описание системы

Основные возможности системы реставрации и обработки изображений старопечатных текстов и рукописей:

1. Загрузка изображений 16 различных форматов (tiff, jpg, bmp, png, gif…)
2. Сохранение изображения в форматах: tiff, bmp, jpg

3. Просмотр исходного и обработанного изображения, а также их фрагментов в произвольном масштабе
4. Возможность одновременной работы с несколькими изображениями
5. Преобразование исходного изображения в полутоновый черно-белый вид
6. Восстановление глобального контраста
7. Адаптивное восстановление локального контраста и яркости по всему изображению
8. Исправление баланса белого
9. Поворот изображения при неточной ориентации страницы во время оцифровки
10. Объединение нескольких фрагментов в одно изображение
11. Адаптивная бинаризация изображения
12. Очистка изображения от шумов и помех
13. Устранение крупных полупрозрачных пятен
14. Устранение проступания надписей с обратной стороны листа
15. Разметка страницы и выявление текстовых блоков и иллюстраций.

Для всех описанных функций обработки изображения существует два режима – полностью автоматический, при котором не требуется никакого вмешательства оператора, и полуавтоматический режим с возможностью ручной коррекции всех параметров. В настоящее время эта система используется в библиотеке Казанского Государственного Университета при создании электронной коллекции газет XIX века.

Описываемая система предназначена для работы на платформе Win9x, Win2k. Время автоматической обработки изображения 2000*1500 пикселей – 15 секунд на компьютере со следующей конфигурацией: PIV-2400, 512 RAM, GeForce-4 MX440 64Mb, WinXP.

Литература

[1] Antique Books Restoration: http://www.units.it/~ipl/research/restoration/antique_books/index.htm

[2] Ramponi G. Digital Automated Restoration of Manuscripts and Antique Printed Books: Доклад на конференции EVA’2005 Florence.

[3] Баженов С.Р., Алексеев В.Н., Бородихин А.Ю., Дергачева-Скоп Е.И., Шабанов А.В. Создание цифровых коллекций редких книг и рукописей из сибирских хранилищ // Тр. конф. “Новые технологии в информац. обесп. науки” - М.: Биоинформсервис, 2001. - С.146-148.

[4] Грузман И.С., Киричук В.С., Косых В.П., Перетягин Г.И., Спектор А.А. Цифровая обработка изображений в информационных системах: Учебное пособие.- Новосибирск: Изд-во НГТУ, 2000. – с.69-73

[5] Масевич А.Ц., Савельев Е.А., Багажков А.К. К созданию электронных коллекций старопечатных книг в библиотеке Российской академии наук: на примере работы над двумя проектами // Тр. конф. “Новые технологии в информац. обесп. науки” - М.: Биоинформсервис, 2001. - С.132-140.

[6] Пратт Э. Цифровая обработка изображений: Пер. с англ,-М.: 1982.—Том.4. с.2-6.

[7] Соловьев В.Д. Электронная коллекция древних книг и рукописей: Исследования по информатике.- Казань: ИПИ АН РТ, 2003. - Выпуск 4. с.21-26.

Об авторах

Южиков В.С. - Казанский Государственный Университет, Казань, РФ
E-mail: Y-Vladimir@yandex.ru