**Структурные методы распознавания сложноорганизованных исторических табличных форм**

В.М. Кляцкин, Е.В. Щепин, К.М. Зингерман, В.В. Лазарев

**Аннотация**

Данная работа посвящена проблематике структурного распознавания сложных табличных форм, встречающихся в исторических источниках. Предложена оригинальная модель "связанных иерархий", в рамках которой может быть распознано и описано обширное семейство табличных форм и бланков. Применение модели "связанных иерархий" в табличном структуризаторе OCR-системы Cript позволило успешно распознавать сложноструктурированные табличные формы из различных исторических источников [1].

**Введение**

Сфера применения систем оптического распознавания непрерывно расширяется не только вширь (по различным областям человеческой деятельности), но и вглубь (по уровню сложности вводимых документов). Уровень сложности зависит от многих факторов и может быть охарактеризован с точки зрения содержания вводимых листов (структура представленной на листе информации, шрифты и другие типографские параметры), качества печати и условий сканирования. Спектр доступных для оптического распознавания структур печатных листов достаточно широк, и тем не менее им не исчерпываются потребности практического использования OCR для распознавания различных исторических источников. В данной работе проведен обзор современных методов структурного распознавания печатной продукции и предпринята попытка расширить диапазон автоматически обрабатываемых OCR- системами структур документов в сторону сложнооорганизованных табличных форм, распространенных в исторических исследованиях. Подлежащие обработке таблицы могут обладать многоуровневой горизонтальной и вертикальной иерархиями, уровни которых могут иметь нерегулярную (не всегда линейно упорядоченную) структуру. Помимо того, структура ячеек таблицы может быть достаточно сложной, несводимой к традиционному описанию текстов "параграф-строка-слово-символ". В ячейках могут встречаться объекты различной структуры, например математические выражения (индексы, дроби, спецсимволы), многоточия, нелинейно-связанные строки. Для анализа подобных структур авторами предложена достаточно универсальная модель "связанных иерархий", позволяющая распознавать, описывать и представлять в текстовой форме широкий класс табличных форм и бланков. Разработанная методология структурного распознавания таблиц основана на модели "связанных иерархий". Программной реализацией этой методологии является табличный структуризатор Cript, входящий в одноименную OCR-систему.

Применение табличной версии OCR-системы Cript к анализу табличных форм из различных исторических источников показало конструктивность предложенной модели и ее алгоритмического наполнения. В статье приведены иллюстрации основных этапов структурного распознавания таблиц исторического характера.

**Обзор методов анализа сложных иерархических форм**

В последние годы системы оптического распознавания символов (OCR-системы) применяются к распознаванию источников различной природы: рукописных текстов, технических публикаций[2], бизнес-карт[3], записей шахматных партий[4] и музыкальных произведений[5], исторических источников[6]. Во многих случаях листы документов, предназначенных для автоматического ввода, определенным образом структурированы. Например, страница книги или научного журнала может быть разбита на колонки, может содержать заголовки, примечания, чертежи и рисунки, страница газет может быть разбита на блоки, содержать оконтуривающие рамки и прочие разделяющие графические элементы.

Одним из наиболее типичных примеров структурированного способа представления информации являются таблицы. Для правильной интерпретации структурированных документов необходимо сначала выделить элементы структуры (колонки и блоки в технических публикациях, графы, строки и ячейки в таблицах) и только после этого распознавать отдельные символы этих элементов[7].

Проблемы выделения структурных блоков на изображении листа документа рассмотрены в ряде работ. Так, создана система выделения блоков текста, основанная на синтаксическом анализе "снизу-вверх" и использующая некоторую информацию о структуре документа, предварительно заданную пользователем [8]. Эта система позволяет выделять на листе изотетичные прямоугольные блоки. При разделении текста на блоки авторы этих работ используют тот факт, что текстовые блоки обычно отделены друг от друга белыми промежутками, ширина которых больше межсловного интервала и интервала между строками. Этот же факт используется для выделения текстовых блоков в других работах[9], но здесь могут иметь более сложную форму. Особенность предложенного в этих работах подхода в том, что он не требует каких-либо предварительных знаний о структуре листа и осуществляет разметку листа, используя только информацию о размерах белых промежутков между габаритными рамками компонент. Метод основан на общем для большинства форматов текстовых документов предположении о том, что текстовые блоки отделяются друг от друга белыми (то есть незаполненными текстом) промежутками. Можно ввести понятие максимальных белых прямоугольников (под которыми понимаются прямоугольники фона, все стороны которых касаются границ рамок компонент текста) и рассматривается алгоритм нахождения всех максимальных белых прямоугольников на растре [10].

Методы, используемые при выделении текстовых блоков, могут быть использованы и для анализа структуры таблиц. Отметим, что во многих случаях (хотя и не всегда) колонки и строки таблиц могут быть разделены линиями разграфки. Используя информацию о расположении этих линий, можно более надежно разделить элементы таблицы, чем используя только информацию о межколонных и межстрочных промежутках в таблице. Однако для этого система анализа структуры листа должна располагать средствами обнаружения линий. Необходимы также средства объединения отрезков линий в прямоугольные рамки, ограничивающие элементы таблиц. Один из алгоритмов быстрого обнаружения горизонталей и вертикалей на изображении и выделения ограниченных ими элементов таблицы представляеьтся весьма эффективным [11]. Приведем краткое описание алгоритма поиска сплошных горизонтальных линий. Растровое бинарное изображение просматривается в направлении сверху вниз и на нем фиксируются последовательности черных пикселов, длина которых превышает заранее заданный порог. Затем связанные (граничащие друг с другом в смысле 8-связности) последовательности объединяются, образуя элементы линий и для каждого такого элемента вычисляется его средняя толщина. Слишком толстые элементы отбрасываются. В качестве кандидатов в элементы пунктирных линий рассматриваются все компоненты связности, которые не слишком велики и являются достаточно "плотными" ("плотность" компоненты определяется как отношение числа черных пикселов к площади ее рамки). Затем для объединения оставшихся компонент в горизонтальные пунктирные линии выполняется следующая процедура: для каждой пары элементов линии проверяется выполнение двух условий: 1)горизонтальный промежуток между двумя элементами линий меньше заранее заданного порога; 2)эти элементы линий перекрываются по вертикали. Если некоторая пара элементов линии удовлетворяет обоим этим условиям, то эти элементы считаются близкими. В результате выполнения этой процедуры выделяются кластеры линейных фрагментов.

Особенность другого алгоритма выделения прямоугольных блоков, ограниченных горизонтальными и вертикальными линиями состоит в том, что сначала посредством скелетизации изображения находятся точки пересечения линий, а затем проверяется наличие самих линий[12]. Средствами выделения текстовых блоков и обработки таблиц располагает также система ExpertVision[1].

\* \* \*

Историками-исследователями за последние годы накоплен определенный опыт по автоматизированному вводу документов сложной сруктуруры в компьютер с использованием различных OCR-пакетов. Так, в ряде работ приведены примеры распознавания отсканированных исторических документов, представляющих собой таблицы различной структуры. В статье Г. Торвальдсена рассмотрены четыре примера распознавания документов, имеющих табличную структуру с помощью OCR-системы Omnipage Professional[14]. На рис. 1a этой статьи приведена копия листа архивной информации, содержащего записи о рождении (отметим, что в таблице отсутствует разграфка), а на рис. 1b - результат распознавания. В работе отмечено, что при обработке документов указанного типа результат может быть интерпретирован в отсутствие информации принадлежности данных к определенной колонке таблицы, поэтому автором не предпринимались какие-либо меры для сохранения этой информации при распознавании. В результате в выходном документе табличная структура данных нарушена. На рис. 2a этой статьи приведено изображение списка эмигрантов. Разделительные линии между колонками и строками в этом исходном документе также отсутствуют. Автор отмечает, что в этом примере формат документа более важен для его понимания, поскольку в одном из столбцов таблицы содержится информация о пункте назначения и стоимости проезда, расположенная в двух смежных строках и относящаяся не к отдельным эмигрантам, а к семьям. Отмечается, что система Omnipage в целом способна правильно распределить информацию по колонкам, вставляя в нужные места строки знаки табуляции, которые могут быть обработаны текстовым процессором. Как видно из рис. 2b, в выходном документе табличная структура входного документа сохранена.

На рис. 3a статьи приведено изображение другого списка эмигрантов, а на рис. 3b - результаты распознавания. На исходном изображении колонки разделены вертикальными линиями, что, по мнению автора, создает дополнительные проблемы для OCR, а именно, способствует перепутыванию символов между колонками. Однако автор считает, что в данном случае эта проблема не является основным препятствием к использованию системы Omnipage. По его мнению, более серьезной проблемой является в данном случае неправильное распознавание отдельных символов. На рис. 4а этой статьи приведено изображение страницы списка налогообложения фермерских хозяйств, а на рис. 4b - результаты обработки этой страницы OCR-системой Omnipage. Можно видеть, что таблица в данном случае имеет сложную иерархическую структуру как по вертикали, так и по горизонтали, причем колонки разделены вертикальными линиями различной толщины и типа (одинарными и двойными). Отмечено, что в данном случае использование системы Omnipage не позволило обеспечить соответствие структуры выходного документа структуре исходной таблицы. Информация часто попадает в несоответствующую колонку, левая колонка опущена. В некоторых случаях (но не всегда) вертикальные линии переносятся в выходной текстовый файл. В результате, по мнению автора, для данного документа эффективнее вводить данные вручную, чем использовать стандартную OCR-систему. Также, по его мнению, можно для автоматизированного чтения этих документов использовать лучшую OCR-систему. В работе Г. Торвальдсена[15] описана дальнейшая обработка документа, который был приведен им ранее [16] в качестве примера. Указано, что преобразование текстового файла, сформированного OCR-системой, в файл, пригодный для ввода в реляционную базу данных, осуществлялось программой, разработанной для этой цели автором. Отмечена ошибка неправильной классификации данных по полям базы.

В работе А. Маквейг рассмотрены вопросы, связанные с автоматизированным чтением таблиц, содержащих статистическую информацию о семейном положении людей в различных районах Ирландии в XIX-XX вв [17]. Обработка изображений выполнялась OCR-системой ProLector. Полученный в результате работы OCR текстовый файл подвергался дополнительной программной обработке с целью выявления ошибок (такая обработка оказалась возможной, поскольку таблицы содержали контрольные суммы), затем эти ошибки исправлялись вручную и осуществлялся ввод исправленной информации в базу данных. Исходные таблицы имеют сложную иерархическую структуру, столбцы этих таблиц имеют различную ширину и разделены вертикальными линиями. Строки, содержащие контрольные суммы, выделены горизонтальными линиями. Автор отмечает, что разделительные линии имели неодинаковую толщину, что приводило к дополнительным затруднениям при их удалении средствами системы ProLector: после удаления линий оставались помехи. Автор указывает также на проблемы, связанные с сохранением структуры таблицы в текстовом файле, сформированном OCR-системой.

В работе М. Олсена рассмотрено применение OCR к обработке списка налогообложения начала XX века в одном из городов провинции Нью-Брунсвик (Канада) [18]. На рис. 2 этой статьи приведено изображение страницы исходного документа, а в табл. 3 - результаты обработки изображения системой Kurzweil (после дополнительной корректировки). Как видно из рис. 2, исходная таблица имеет простую структуру, колонки ее разделены вертикальными линиями, имеющими многочисленные разрывы. Автор отмечает, что OCR-система Kurzweil испытывает некоторые трудности при обработке вертикальных линий, вследствие чего возникают ошибки, связанные с неправильным размещением данных в колонках.

Авторы всех вышеупомянутых работ по применению OCR при выполнении исторических проектов отмечают, что использование OCR-систем для автоматизированного ввода в ЭВМ документов, содержащих таблицы, позволило значительно сэкономить время по сравнению с ручным вводом данных в ЭВМ (перенабивкой). Вместе с тем, значительные усилия потребовались для поиска и исправления ошибок. Одной из причин появления ошибок явилось отсутствие достаточно мощных средств описания и распознавания структуры таблиц в использованных OCR- системах.

**Модели описания структуры текстовых и табличных документов**

Модель описания структуры текстовых документов - дерево регулярного ортогонального чередования разбиений.

Традиционной для текстовых документов является следующая древовидная структура листа: лист разбит на текстовые блоки, называемые колонками, колонки разбиты на параграфы, параграфы состоят из слов, слова - из символов, символы - из одной или нескольких компонент связности.

Используемая в системе CRIPT структура описания листа представляет собой обобщение традиционной древовидной схемы. Существенными являются следующие особенности реального графа описания формата листа:

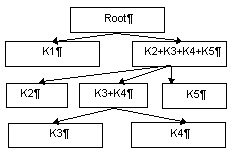


Рис. 2. Дерево колонок, соответствующее структуре листа, схематически изображенного на рис. 1a.

Для каждого нетерминального объекта (т.е. объекта, имеющего "дочерей"), сохраняется информация об относительном взаимном расположении его дочерей на листе, что позволяет в случае необходимости сохранить в выходном документе то же взаимное расположение структурных элементов текста в пределах иерархического блока более высокого уровня, что и в исходном документе. Предусмотрены, в частности, следующие варианты взаимного расположения объектов: упорядочение по горизонтали, упорядочение по вертикали, матричный порядок, индексный порядок (одна из дочерей является индексом другой), неупорядоченное расположение.

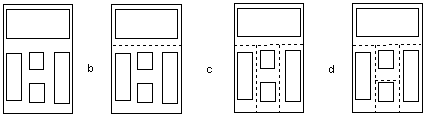


Рис.1 Последовательные шаги расщепления листа на колонки по методу "сверху вниз"

Колонки образуют иерархическую структуру, т.е. каждая колонка может быть разбита на подколонки, причем на соседних уровнях ориентация разделяющих колонки линий меняется на ортогональную (так, например, если на некотором уровне колонки могут быть разделены горизонталями, то на следующем уровне они разделяются вертикалями). Именно поэтому такая модель описания структуры документов была названа деревом регулярного ортогонального чередования разбиений. На рис. 1 приведена схема, поясняющая данную модель. На рис. 1,a схематически показано расположение колонок на листе, на рис. 1,b-d - последовательные шаги расщепления, на рис. 2 - построенное "дерево регулярного ортогонального чередования разбиений"

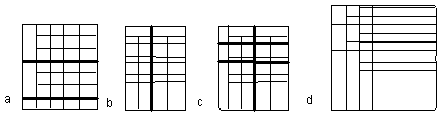


Рис. 3. Примеры таблиц иерархической структуры: a) Таблица с горизонтальной иерархией; b) Таблица с вертикальной иерархией; c) Таблица с горизонтальной и вертикальной иерархией; d) Таблица с двумя уровнями горизонтальной иерархии.

Строки могут быть разбиты на подстроки, различающиеся какими-то свойствами (например, шрифтом).

Объекты иерархической структуры могут иметь дочерей как на следующем по порядку иерархическом уровне (например, дочерьми параграфа могут быть строки), так и на более низком иерархическом уровне (например, дочерьми колонки могут быть символы).

Модель описания структуры табличных форм - обобщенная модель связанных ортогональных иерархий;

Рассмотрим особенности табличных структур, определяющие необходимость использования специальной модели для их описания:

Правильное понимание таблицы читателем невозможно без учета информации о взаимном расположении строк, колонок и ячеек таблицы. Поэтому при автоматизированном распознавании табличных форм необходимо в выходном документе сохранить то же взаимное расположение этих структурных табличных элементов, что и в исходной таблице.

Строки и колонки таблиц могут иметь иерархическую структуру (рис. 3), причем иерархия может быть многоуровневой (рис. 3,d).

Колонки и горизонтальные блоки различных иерархических уровней в таблицах могут иметь заголовки. Ориентация текста в заголовках может быть как горизонтальной, так и вертикальной.

Колонки и строки в таблицах могут быть разделены линиями. Толщина линии, ее тип (например, одинарная или двойная, сплошная или пунктирная), а также сам факт наличия или отсутствия разделительной линии между двумя колонками или строками таблицы могут быть связаны определенным образом со структурой и семантикой таблицы и, следовательно, при автоматизированном распознавании таблиц должны быть отражены в случае необходимости в выходном документе.

С учетом перечисленных особенностей ясно, что ни дерево регулярного ортогонального чередования, ни матричная структура в большинстве случаев не подходят для моделирования структуры табличных форм. Хотя дерево регулярного ортогонального чередования и сохраняет некоторую информацию о взаимном расположении элементов таблицы, этой информации недостаточно, чтобы воспроизвести исходную структуру таблицы в выходном документе. Пример приведен на рис. 4. Хотя структура таблиц на рис. 4, a) и b), различна, деревья регулярного ортогонального чередования, описывающие структуру этих таблиц, совпадают (см. рис. 4c). Неприменима в общем случае к таблицам и матричная структура, так как она не позволяет описывать таблицы с иерархической организацией.

Чтобы выяснить, какая модель данных может быть использована для представления таблиц, рассмотрим сначала общепринятую в технической документации структуру таблиц. Обычно таблица состоит из шапки, может быть, боковика и так называемого "тела" таблицы (рис. 5), причем как шапка, так и боковик могут иметь иерархическую структуру (рис. 6). Модель такой таблицы может состоять из дерева горизонтальной иерархии, описывающего структуру боковика, дерева вертикальной иерархии, описывающего структуру шапки и связей между деревьями горизонтальной и вертикальной иерархий.

Рассмотренную модель будем называть обобщенной моделью связанных ортогональных иерархий. Пример описания структуры таблицы с использованием этой модели приведен на рис. 7 (b,c).

**Общий алгоритм распознавания таблиц**

Система распознавания табличных форм состоит из следующих основных компонент: 1) подсистемы предобработки; 2) табличного структуризатора; 3) подсистемы распознавания символов (OCR-систе-мы); 4) табличного форматизатора; 5) подсистемы постобработки.

OCR-система осуществляет распознавание текстовой информации на изображении (основа метода распознавания рассмотрена в работе[19].

Табличный форматизатор выводит таблицу, структура которой определена структуризатором, в виде текстового файла, располагая распознанные OCR-системой символы в надлежащем порядке в соответствующих ячейках таблицы. Подсистема постобработки осуществляет контекстно- зависимую коррекцию символов (например, заменяет при необходимости строчные буквы на прописные в зависимости от расположения их в слове).

Рассмотрим более подробно работу табличного структуризатора. Эта подсистема последовательно выполняет следующие действия: 1)обнаружение и удаление линий на растре; 2)автономную обработку линий; 3)автономную обработку текста; 4)распознавание структуры таблиц; 5)анализ ячеек таблицы.

Процедура обнаружения линий обрабатывает растровое изображение листа документа и обнаруживает на нем отрезки прямых линий. Затем эти линии удаляются с растрового изображения, чтобы отделить их от изображений символов и обеспечить возможность автономной обработки линий и текста.

Процедура автономной обработки линий выполняет следующие действия: а)сращивание разрывов линий; б)оценивание и ликвидация перекоса; в)выделение кластеров соосных линий; г)определение типов линий. Процедура автономной обработки текста состоит в основном в контекстно-независимой фильтрации обрывков линий и фильтраци шумов на текстовом слое растра. Входными данными для процедуры распознавания табличной структуры являются: априорное описание структуры таблицы; набор отрезков горизонтальных и вертикальных линий, классифицированных по типам; информация о положении и размерах рамок текстовых компонент связности.

Эта процедура проверяет, соответствует ли структура обрабатываемого изображения априорному описанию, и в случае соответствия формирует иерархическую модель структуры таблицы. Для того, чтобы обеспечить устойчивое распознавание, эта процедура учитывает специфические особенности таблиц определенного типа. Основные действия, выполняемые этой процедурой:

идентификация горизонтальных линий, ограничивающих шапку;

глобальная контекстно-зависимая фильтрация линий;

идентификация вертикальных линий в соответствии с априорным описанием и проверка соответствия структуры таблицы этому описанию;

контекстно-зависимая фильтрация компонент (обрывков линий и шума);

построение вертикальной иерархии (структуры колонок);

оценивание горизонтальной иерархии.

Назначение процедуры анализа ячеек - структуризация текста внутри ячеек таблицы. Процедура выполняет следующие действия: выделение строк и слов, обнаружение знаков препинания, индексных выражений, дробей и многоточий внутри ячеек таблицы.

**Заключение**

Таким образом, в настоящей работе проведен обзор современных технологий автоматизированного ввода в компьютер сложноструктурированных печатных документов и сделан вывод о наличии определенной ограниченности средств структурного распознавания таблиц сложной формы.

Авторами предложена достаточно универсальная структурная модель связанных иерархий, пригодная для описания широкого класса текстовых документов, табличных форм и бланков. В работе описывается основанная на модели связанных иерархий методология распознавания табличных структур. Данная методология была положена в основу разработанного в рамках OCR-системы Cript универсального текстово-табличного структуризатора.

Приведенные в работе базовые алгоритмы структуризатора Cript могут найти применение во многих задачах автоматизированной обработки текстов и таблиц различной природы, что демонстрируется примерами структурного распознавания таблиц из различных исторических источников. Поэтому погружение OCR системы Cript в различные автоматизированные системы ввода и обработки исторических источников позволит повысить эффективность и степень автоматизации подобных систем. Гибкость и простота разработанного в рамках проекта Cript языка описания табличных форм существенно упростят процедуру настройки конечным пользователем OCR-системы на специфические особенности вводимых табличных структур.

Авторы выражают признательность А.В. Тюленеву и Л.И. Бородкину, благодаря которым во многом и состоялась данная статья.

**Список литературы**

1.OCR (Optical Character Recognition) - оптическое распознавание образов, Cript - одна из эффективных отечественных OCR-систем (Прим. ред.)

2.L. O'Gorman, "The document spectrum for page layout analysis", IEEE Journal on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, p. 1162-1173 ,1993; M. Krishnamoorthy, G. Nagy, S. Seth, M. Viswanathan,"Syntactic segmentation and labelling of digitized pages from technical journals", IEEE Journal on Pattern Analysis and Machine Intelligence, vol.15, no.7, p.737-747, 1993; G. Nagy, S. Seth, M. Viswanathan, A prototype document image analysis system for technical journals, Computer, no.7, p.10-21, 1992.

3.L. O'Gorman, The document spectrum for page layout analysis, IEEE Journal on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, p. 1162-1173 ,1993

4.H.S. Baird, K. Thompson, Reading Chess, Proc. of Workshop on ComputerVision, Washington, 1987.

5.V. Poulian d'Andecy, J. Camillerapp, I. Lemplumey, Kalman Filtering for Segment Detection: Application to Music Scores Analysis, Proc. IAPR 12-th Intern. Conf. on Pattern Recognition, Los Alamos,1994. - v.1.

6.M. Thaller. The Processing of Manuscripts, Images and Manuscripts in Historical Computing, Goettingen, 1992; T. Kruyt, J. van der Voort van der Kleij. Converting the Historical Dictionary of Dutch to Electronic Form, Optical Character Recognitionin The Historical Discipline., Goettingen, 1993; 22. L. Borodkin, V. Lazarev, E. Zlobin, Applications of OCR in Russian Historical Sources: a Comparison of various programs, Optical Character Recognition in The Historical Discipline, Goettingen, 1993.

7.V. Klyahzkin, E. Shchepin, K. Zingerman. Application of hierarchical methods of cluster analysis to the printedtext structure recognition, Shape, Structure, and Pattern Recognition, Dov Dori and Alfred Bruckstein, Eds. World Scientific, 1995; V. Klyahzkin, E. Shchepin, K. Zingerman. Hierarchical analysis of multi-column texts, Pattern Recognition and Image Analysis, Vol.5, No.1, 1995, Interperiodica,pp. 1-12; В.М. Кляцкин. Иерархический кластер- анализ многоколонных текстов, Одесса, 1994, Труды V Международной конференции (Статистический и дискретный анализ данных и экспертные оценки), Изд-во Одесского политехнического университета, стр. 132-134.

8.M. Krishnamoorthy, G. Nagy, S. Seth, M. Viswanathan, Syntactic segmentation and labelling of digitized pages from technical journals, IEEE Journal on Pattern Analysis and Machine Intelligence, vol.15, no.7, p.737-747, 1993

9.H.S. Baird, S.E. Jones, S.J. Fortune, Image Segmentation by Shape-Directed Covers, Proc. IAPR 10-th Intern. Conf. on Pattern Recognition, Atlantic City, 1990; H.S.Baird, Background Structure in Document Images, Proc. IAPR Workshop on Structural and Syntactic Pattern Recognition, Bern, 1992.

10.H.S. Baird, Background Structure in Document Images, Proc. IAPR Workshop on Structural and Syntactic Pattern Recognition, Bern, 1992.

11.H. Saiga, Y. Kitumura, S. Ida, "High-Speed Recognition of Tabulated Data", Proc. IAPR 12-th Intern. Conf. on Pattern Recognition, Los Alamos, 1994. - v.2.

12.J. F.Arias, A. Prasad, R. Kasturi, A.Chhabra, Interpretation of Telephone Company Central Office Equipment Drawings, Proc. IAPR 12-th Intern. Conf. on Pattern Recognition, Los Alamos, 1994. - v.2.

13.P. Wayner, Optimal Character Recognition , Byte, 1993, no. 12, p.203-210.

14.G. Thorvaldsen. Making Printed Historical Sources Machine Readable: Some Experiences with OCR, History and Computing, Vol.5 No 2,Edinburgh University Press, 1993.

15.G. Thorvaldsen. Historical OCR Projects in Norway, Optical Character Recognition in The Historical Discipline., Goettingen, 1993

16.G. Thorvaldsen. Making Printed Historical Sources Machine Readable: Some Experiences with OCR, History and Computing, Vol.5 No 2,Edinburgh University Press, 1993.

17.A. McVeigh. The Irish Database Project: A Case for OCR?, Optical Character Recognition in The Historical Discipline., Goettingen, 1993

18.M. Olsen. Scanning, Keyboarding and Data Verification: Factors in Selecting Data Collection Technologies, Optical Character Recognition in The Historical Discipline., Goettingen, 1993

19.E. Shchepin, G. Nepomnyashchii, Character recognition via critical points, International Journal of Imaging Systems and Technology, vol.3, pp.213-221, 1991

20.Материалы для географiи и статистики Россiи, собранные офицерами Генеральнаго штаба. Казанская губернiя. (составитель М. Лаптев) - Санкт-Петербург, Военная типография. - 1861.