FrameStream: модель сегментации многостраничного структурированного документа

В. В. Постников, А. Е. Марченко

ООО «Когнитивные технологии»,
Россия, 117312 Москва, пр. 60-летия Октября, 9

В работе рассматривается центральная часть задачи ICR (Intelligent character recognition) — этап сегментации, состоящий в идентификации на графических образах частей документа, указанных в шаблоне документа. Постановка задачи допускает нежесткое разделение документа на страницы и свободное перетекание текстовых блоков. Предлагается модель, отражающая состояние процесса сегментации как однородной структуры вложенных фреймовых потоков. Структура кодируется графом специального вида, который наращивается в процессе сегментации и распознавания. Верхний уровень модели соответствует потоку отсканированных страниц, нижний уровень — потоку символов в текстовой строке. Модель допускает альтернативные варианты декомпозиции и распознавания, а также присутствие разных типов документов в одном потоке и разных вариантов форматирования документов. В такой парадигме задача распознавания сводится к построению и поиску наилучшего (с точки зрения соответствия модели документа) пути в графе сегментации.

Введение

Распознавание многостраничных структурированных документов — актуальная, но нерешенная задача. Существующие OCR-системы успешно распознают многостраничные документы (книги, журналы, отчеты и т. п.).

В процессе работы OCR-система автоматически выделяет строки (фрагменты строк), распознает строки, компонует строки в абзацы и текстовые колонки. Как результат, основной единицей на выходе OCR-системы является текстовый абзац или фрагмент абзаца (если он оказался перенесенным на другую страницу или колонку). Система пытается отформатировать полученный набор абзацев таким образом, чтобы он визуально был похож на исходный документ. Это приемлемо, если мы хотим редактировать документ в текстовом процессоре или индексировать поисковыми машинами в Интернете.

ICR-системы, в отличие от OCR-систем, распознают отсканированный документ в контексте шаблона (формы) документа. Шаблон специфицирует логическую и визуальную модель документа. С точки зрения логической модели документ представляет собой информационный объект известной структуры (т. е. состоящий из известного набора частей, по определенным правилам скомпонованных). Визуальная модель определяет, каким образом логические части *представлены на бумаге* (синтаксис и стиль заполнения, правила форматирования, взаимное расположение и т. п.). В процессе обработки документа ICR-система пытается выделить значимые части на обрабатываемом документе, соотнести их с шаблоном, распознать и, как результат, — создать информационный объект (запись в базе данных, xml-файл или т. п.), который может в дальнейшем полноценно обрабатываться на компьютере.

На сегодняшний день существует немало стандартных форм документов, которые успешно обрабатываются ICR-системами, — это банковские платежные поручения, разного рода анкеты и т. п. Перечисленные в [1–8] и [10, 11] работы рассматривают разные аспекты этой «одностраничной» задачи ICR-распознавания. Однако существует также гораздо более широкий класс документов, которые могли бы, но не могут (или могут лишь частично) на сегодняшний день обрабатываться ICR-системами. К ним относятся многостраничные документы, в которых модель содержания известна заранее, но текстовые части «перетекают» со страницы на страницу, из колонки в колонку и т. д. Это исключает возможность постраничного разбора документа, на который ориентированы существующие ICR-системы. Разработке модели сегментации такого вида документов посвящена данная работа.

1. Содержательное описание модели

Модель FrameStream кодирует состояние процесса сегментации документа (в задаче распознавания под управлением шаблона). Отметим, что модель не привязана к формализму (языку, схеме), в котором представлен шаблон документа. В приведенном ниже примере шаблон документа опи-

сан вербально, в стиле того, как это часто делают издатели журналов. (Для тестирования модели авторы использовали систему Cognitive Forms, в которой описание модели документа производится средствами CFML — Cognitive Forms Markup Language [9].)

В предлагаемой модели документ декомпозируется на структуру вложенных фреймовых потоков. Под фреймом понимается прямоугольник графического образа. Последовательность фреймов образует так называемый фильм. Изначально документ представлен фильмом, состоящим из фреймов-страниц. Если процесс сегментации документа идет сверху вниз (от общего к частному), фильмы детализируются и из них выделяются вложенные фильмы. В варианте сегментации снизу вверх низкоуровневые фильмы могут порождать фильмы более высокого уровня (например, из букв могут собираться строки, из строк абзацы и т. д.). Конечная цель разбора состоит в установлении соответствия фильмов (фрагментов графического образа отсканированных страниц) логическим частям документа.

Состояние процесса сегментации документа кодирует орграф специального вида (граф сегментации). Помимо вершин-фреймов граф содержит вершины-разрезы двух видов (начало и конец фрейма) и вершины-хабы (начало и конец фильма), играющие роль заголовков и окончаний вложенных списков. Дуги кодируют отношение порядка на фреймах и разрезах, а также вложенность фильма в фильм. Интерпретация дуг однозначно определяется типом их начальной и конечной вершин. На рис. 1 приведен граф, соответствующий верхнему уровню сегментации документа, представленного как последовательность фреймов-страниц Page1 и Page2.

В процессе разбора граф сегментации наращивается. В процессе сегментации могут возникать альтернативные варианты разбора, что может быть связано с неоднозначностью процесса распознавания — разными вариантами проведения разрезов, многозначными ответами модулей распознавания символов, а также с различными вариантами, допустимыми в рамках модели документа. Модель позволяет кодировать альтернативные способы сегментации, насыщая граф дополнительными вариантами. В конечном счете результат сегментации документа представляется как путь в построенном графе, обеспечивающий наиболее правдоподобную (соответствующую модели документа) сегментацию. В случае если построенный путь имеет низкую оценку качества, фрагменты графа могут быть декомпозированы альтернативным способом, что приведет к появлению дополнительных фрагментов путей в графе и может улучшить оценку. Однородность модели позволяет производить построение альтернативных вариантов «широким фронтом», параллельно на всех уровнях декомпозиции.

2. Пример сегментации документа в модели FrameStream

Рассмотрим в качестве примера процесс сегментации документа «научная статья». Как правило, статьи в научном журнале или сборнике содержат одинаковые части и однородно отформатированы. Представим, что необходимо отсканировать журнал (или подписку за последние годы) и сформировать базу данных статей. Логически статья содержит части: Title (название), Authors (список авторов), Abstract (аннотация), Text (основной текст), RefList (список литературы). Эти части могут быть декомпозированы более подробно, например в списке литературы могут быть выделены элементы списка, а в элементах списка — название статьи, ее авторы и т. д. Физически статья — это набор страниц. Возможна ситуация, когда статья в журнале начинается на продолжении страницы, на которой закончилась предыдущая статья. В таком случае сборник или его раздел начинается с новой страницы и модель должна быть построена на уровне сборника или раздела.

Для простоты изложения мы ограничимся уровнем статьи, считая, что каждая статья сборника начинается с новой страницы. Введем правила форматирования — заголовок, под ним располагается список авторов, затем текст разбивается на две колонки, в которых последовательно расположены аннотация, основной текст и список литературы. На рис. 2 представлена схема текстовых блоков одной из статей. Приведенный пример содержит восемь блоков — Title, Authors, Abstract, Text1, Text2, Text3, Text4 и RefList. Блоки расположены на страницах Page1 и Page2.

Правила форматирования статьи, закодированные в шаблоне документа, указывают, что блоки Title и Authors располагаются в начале документа и могут быть отделены от остального текста горизонтальным разрезом. Первый шаг сегментации — преобразование фильма {Page1, Page2} в фильм {Frame1, Frame2, Frame3} (см. рис. 3, 4).

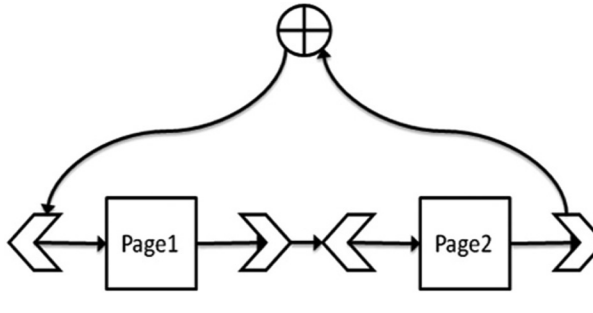

Рис. 1

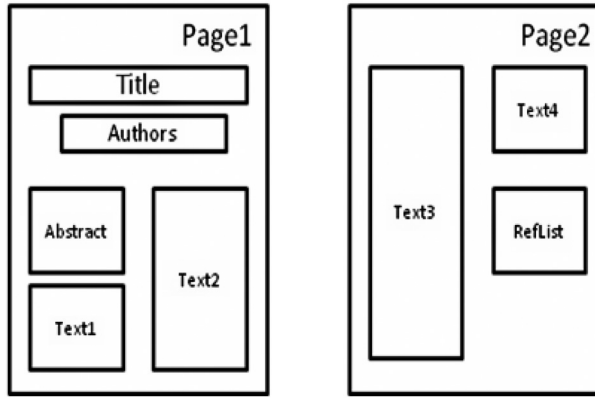

Рис. 2

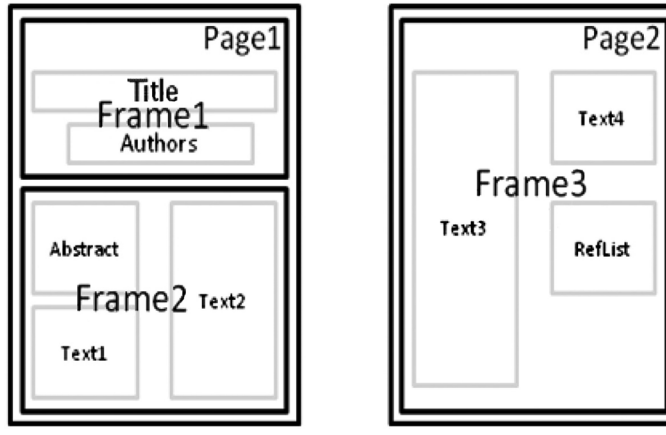

Рис. 3

В приведенной схематической нотации вершины графа сегментации, соответствующие фреймам, обозначены прямоугольниками, вершины-разрезы — угловыми скобками («начало фрейма» и «конец фрейма»). Начало и конец фильма объединены на схеме в вершину-хаб, которая обозначается окружностью с перекрестием. Таким образом, хаб представляет собой кортеж из двух вершин: ⟨«начало фильма», «конец фильма»⟩. Каждая из них имеет не более одной входящей и не более одной исходящей дуги. Вершина «начало фильма» соединяется только с вершинами типа «начало фрейма»; «конец фильма» — только с вершинами типа «конец фрейма». Это позволяет упростить визуальную нотацию, не потеряв однозначности определения путей обхода. Трактовка путей обхода представлена на рис. 5.

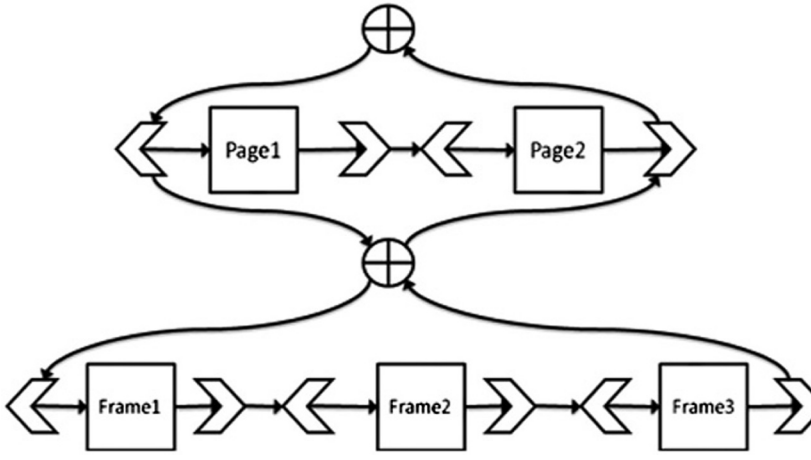

Рис. 4

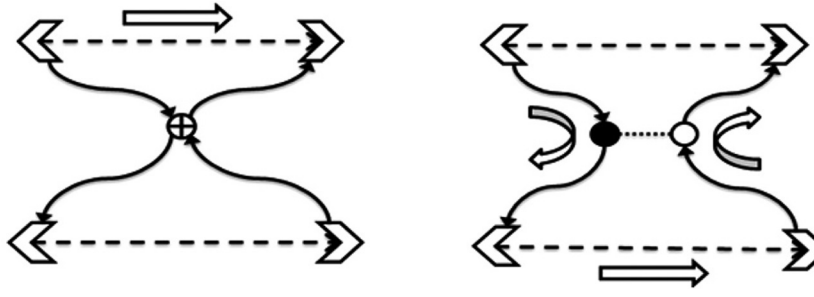

Рис. 5

На следующем шаге сегментации в приведенном примере фрейм Frame1 (заголовок статьи) декомпозируется на фреймы Title и Authors (рис. 6). Отметим, что выделенные фреймы Title и Authors уже соответствуют нижнему уровню описания документа и позволяют не только распознать текст в указанных зонах документа, но и разместить полученные результаты в соответствующих полях базы данных, т. е. как «название» и «список авторов» распознанной статьи. Поскольку в общем случае текст может содержаться в нескольких последовательных фреймах («перетекать»), блоки Title и Autors замыкаются в графе как отдельные фильмы.

Согласно модели документа, его оставшаяся часть содержит текст, отформатированный в две колонки. Соответственно, фильм {Frame1, Frame2} преобразуется в фильм {Frame2_1, Frame2_2, Frame3_1, Frame3_2}. Изменение этой части графа в процессе данного шага отражено на рис. 7.

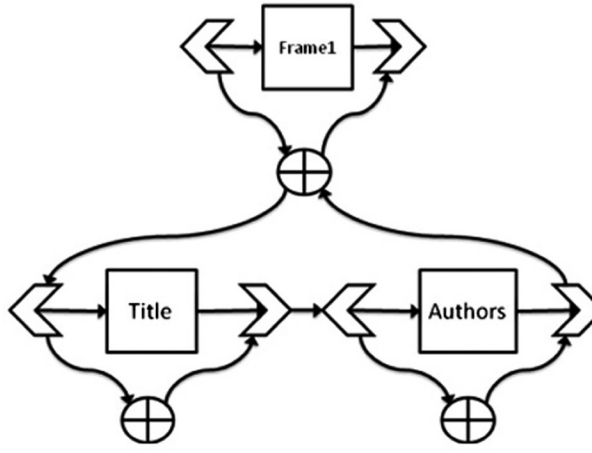

Рис. 6

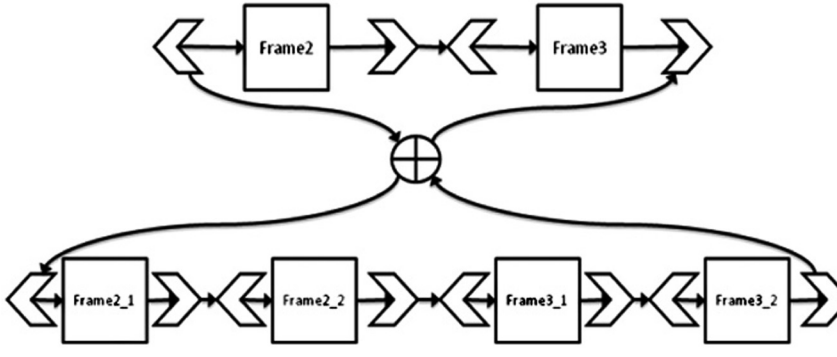

Рис. 7

Далее фильм {Frame2_1, Frame2_2, Frame3_1, Frame3_2} декомпозируется горизонтальными разрезами на фильм {Abstract, Text1, Text2, Text3, Text4, RefList}, из которого выделяются фильмы {Abstract}, {Text1, Text2, Text3, Text4}, {RefList}. Эти три фильма соответствуют частям логической модели документа. Состояние графа, соответствующее данному этапу, отражено на рис. 8. Пунктирные стрелки отражают отношение вложенности фрейма во фрейм.

Отметим, что к данному этапу сегментации хабы нижнего уровня соответствуют частям логической модели документа. Процесс декомпозиции может быть продолжен. Соответствующие выделенным фильмам текстовые блоки могут быть далее декомпозированы на строки, слова и буквы. В резуль-

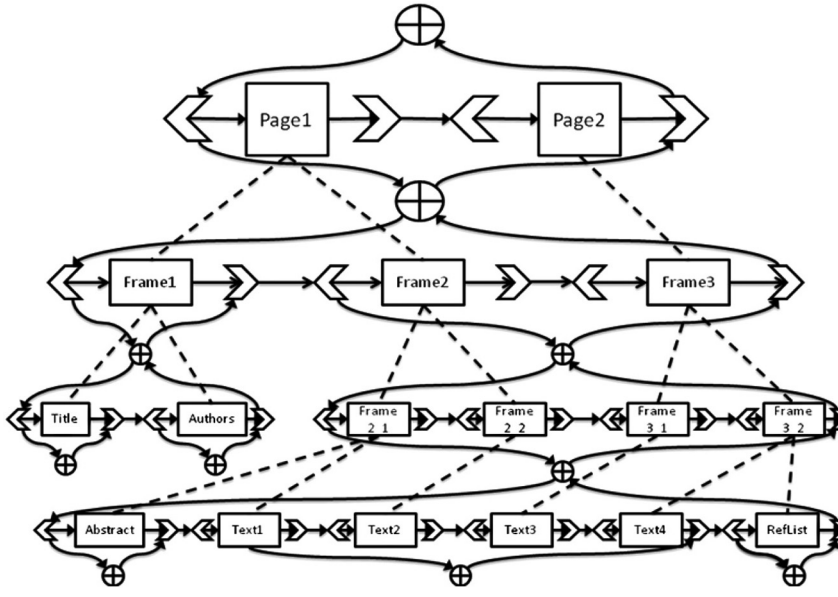

Рис. 8

Дата выдачи 22.05.06 13:58

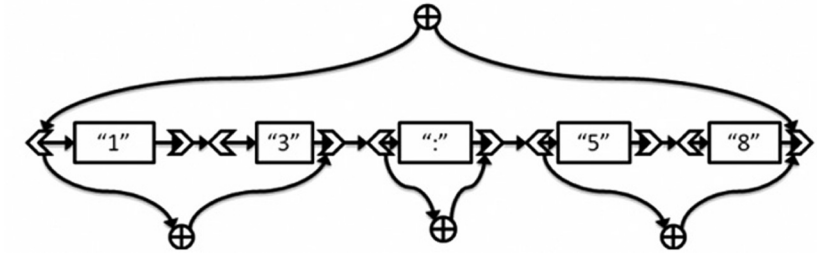

Рис. 9

тате графический образ текста сопоставляется с логической частью документа и должным образом интерпретируется.

На рис. 9 приведен пример терминального этапа разбора, иллюстрирующий декомпозицию фрагмента документа «время выдачи».

3. Кодирование альтернативных вариантов сегментации

В процессе автоматической сегментации документа могут возникать альтернативные варианты сегментации. Существует две основные причины, приводящие к неоднозначности при разборе документа.

1. Собственно модель документа может допускать разные варианты форматирования. В том числе, входящий поток документов может содержать разные виды документов. Продолжая рассмотренный пример с документом «научная статья», можно предположить, что поток распознаваемых документов может содержать два типовых варианта форматирования (см. рис. 10 и 11).

2. Процесс идентификации и распознавания элементов документа может иметь многозначные ответы. Могут существовать альтернативные варианты выделения на графическом образе фрагментов (областей пикселей), относящихся к частям документа — колонкам, абзацам, строкам, словам, символам. К неоднозначности сегментации добавляется неоднозначность интерпретации выделенных фрагментов.

Альтернативные варианты сегментации порождают разветвления в графе сегментации, преобразуя фильм как локальную цепочку фреймов в сеть, каждый путь в которой является альтернативным вариантом фильма. На рис. 12 и 13 показан пример, когда фильм {A, B} должен (из требования модели) быть разделен на два подфильма. Но проведение разреза может быть неоднозначно. На рис. 12 приведен вариант, когда фрейм A может быть разделен двумя способами. В результате образуется две пары подфильмов $P1 = \langle F1 = \{A11\}, F2 = \{A21, B1\}\rangle$ и $P2 = \langle G1 = \{A12, B1\}, G2 = \{A22, B1\}\rangle$.

Альтернативные варианты разреза не обязательно должны присутствовать в рамках одного фрейма. На рис. 13 приведен пример графа сегментации, когда один вариант разреза разделяет фрейм A, а другой (альтернативный) разделяет фрейм B.

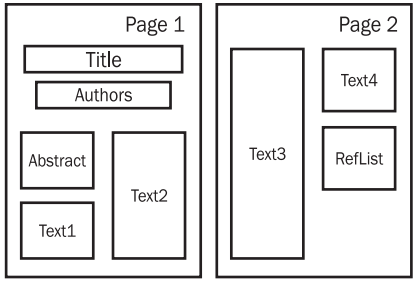

Рис. 10

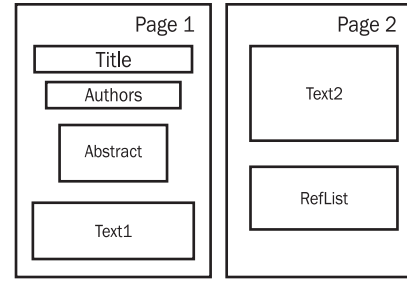

Рис. 11

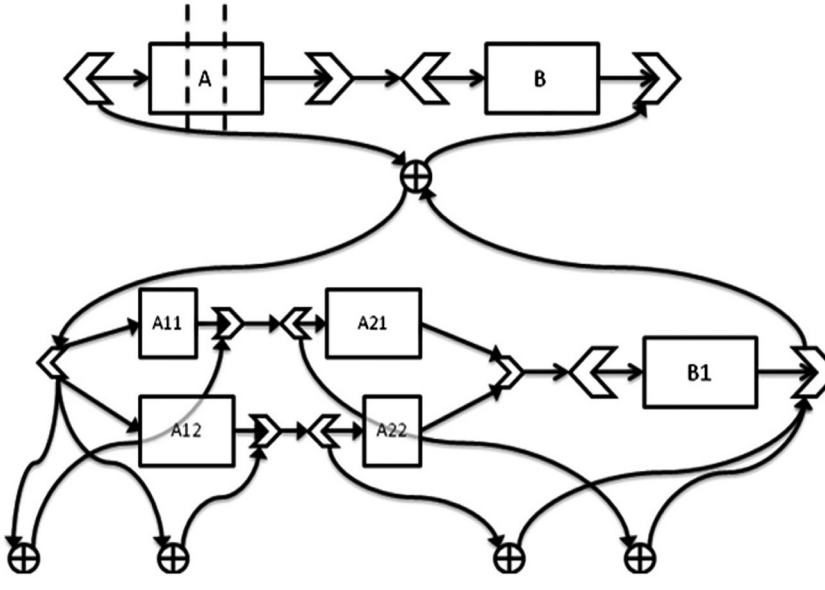

Рис. 12

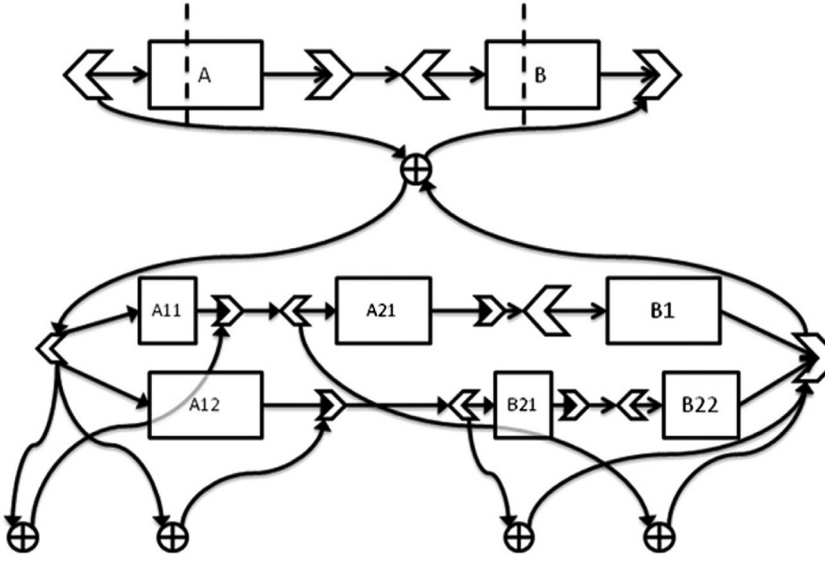

Рис. 13

Литература

1. *Постников В. В.* Разработка методов наложения формы на графическое изображение документа // В сб. Интеллектуальные технологии ввода и обработки информации. М.: URSS, 1998.
2. *Postnikov V. V.* Flexible Forms Identification. Proceedings of the 5th German-Russian Workshop on Pattern Recognition and Image Understanding (GRWS98). Hamburg: Infix, 1999.
3. *Постников В. В.* Формальный подход к задаче идентификации графических образов структурированных документов // В сб. Развитие безбумажных технологий в организационных системах. М.: URSS, 1999.
4. *Арлазаров В. В., Постников В. В., Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Управление информационными потоками. М.: URSS, 2002.
5. *Postnikov V. V.* Identification and Recognition of Documents with a Predefined Structure // Pattern Recognition and Image Analysis. 2003. Vol. 13. № 2. P. 332–334.
6. *Postnikov V. V., Sholomov D. L., Marchenko A. E.* FlexiDocs: The Template Driven Document Recognition Technology. Proceedings of the 6th German-Russian Workshop on Pattern Recognition and Image Understanding (OGRW-6), 2003.
7. *Postnikov V. V., Sholomov D. L.* Post-processing of OCR Results Using Automatically Constructed Partially Defined Syntax // Proc. of The International Conference on Machine Learning, Technologies and Applications. USA: CSREA Press, June 2004.
8. *Постников В. В., Марченко А. Е., Шоломов Д. Л.* Разбор структурированного документа в модели с нечеткой логикой // Документооборот. Концепции и инструментарий. М.: URSS, 2004.
9. *Постников В. В., Марченко А. Е.* CFML: язык описания многостраничных структурированных документов для их идентификации и распознавания. Математические методы распознавания образов (ММРО-12): Сборник докладов 12-й Всероссийской конференции. М.: МАКС Пресс, 2005.
10. *Усилин С. А., Николаев Д. П., Постников В. В.* Быстрый алгоритм совмещения изображений документов в произвольной геометрической модели // Труды конференции «Информационные технологии и системы» (ИТиС). Геленджик, 2008. С. 471–477.
11. *Безматерных П. В., Николаев Д. П., Постников В. В.,* Метод идентификации типа документа по структуре проекций его изображения на координатные оси // Труды конференции «Информационные технологии и системы». (ИТиС). Геленджик, 2008. С. 498–501.