Система разбора документа, заданного атрибутами структурных элементов и отношениями между структурными элементами*

А.Е. Марченко, Е.И. Ершов, С.А. Гладилин

Аннотация. В рамках задачи распознавания документов при помощи технологий компьютерного зрения рассматривается задача сопоставления структурных элементов документа с их физическими образами на бумаге, при условии, что элементы не имеют фиксированного расположения. Предлагается подход, основанный на описании документа через атрибуты его структурных элементов и отношения между структурными элементами. Предлагается алгоритм разбора документа, использующий данный подход. Описывается реализованная система разбора документа, основанная на данном подходе.

Ключевые слова: *разбор документа, структурный элемент, отношения между элементами, атрибуты элементов, алгоритм разбора.*

Введение

Задача компьютерного зрения по распознаванию документов, структурные элементы (СЭ) которых не имеют фиксированного расположения на бумаге, приобретает все большую важность по мере развития электронного документооборота.

Отсутствие фиксированного расположения не подразумевает полного отсутствия ограничений на местоположение СЭ. Для разного характера и степени жесткости этих ограничений предложены различные методы разбора документа. Случай, когда вариации в расположении СЭ настолько малы, что возможно растровое наложение распознаваемого документа и некоего эталонного образа, рассмотрены в [1] и [2].

Для менее жестких ограничений производится отход от растрового наложения целого документа в сторону сопоставления отдельных СЭ документа с их образами на бумаге. Для этого нужна декомпозиция документа на набор составляющих объектов с указанием связей между ними. Обобщение данного подхода позволяет перейти к методам распознавания также других объектов, заданных формальным описанием наблюдаемых свойств. Общим в этих методах будет декомпозиция распознаваемой сущности на СЭ с указанными свойствами и задание графа связей между СЭ.

В существующих подходах [3, 4, 5] к распознаванию документа его СЭ считают текстовые строки, линии разграфки и другие элементы, встречающиеся в большинстве или значительном числе документов. Для распознавания СЭ в упомянутых методах используются отдельные модули. В настоящей работе эти модули будут называться первичными детекторами.

При разборе документов в большинстве случаев используются первичные детекторы текстовых строк. Таковыми детекторами чаще всего выступают модули оптического распознавания текста [6]. Существует также подход к детекции строк, основанный на алгоритме распознавания объектов. Для строк с фиксированным текстом и жестко заданными шрифтовыми атрибутами данный подход демонстрирует значительно более высокую надежность [7].

Под моделью представления документа будет пониматься совокупность описаний СЭ, их свойств (атрибутов) и связей между СЭ. Модель представления является входными данными системы распознавания, наряду с образом документа. Системы распознавания документов различаются гибкостью и универсальностью моделей представления, а также способами их задания.

В ряде систем на модель представления документа наложены существенные ограничения, сокращающие множество видов связей между СЭ. Это упрощает как построение модели, так и сам процесс разбора документа, однако при этом уменьшает универсальность системы. Таковы, например, системы распознавания документов табличной структуры [8].

Особой ветвью задачи разбора документа является разбор документов свободной структуры,

* За счет средств гранта РНФ (проект #14-50-00150) предложен подход распознавания, основанный на описании объекта через свойства его элементов, а также предложен алгоритм разбора документа с целью распознавания, реализующий данный подход.

т.е. разбор в отсутствие модели представления. Такой подход вычленяет из распознанного документа СЭ и отношения между ними на основании только лишь самой общей информации о возможно присутствующих СЭ, их свойствах и допустимых отношениях [9]. Часть систем, реализующих этот подход, опирается только на графическое изображение документа, не используя предварительное распознавание текстовых строк [10], другие опираются как на графический образ, так и на результаты предварительного распознавания [11], третьи используют двухэтапный подход, при котором стадия предварительного разбора происходит без использования результатов текстового распознавания, результат предварительного разбора используется для распознавания строк, и окончательный разбор выполняется уже с распознанными строками [12]. Недостатком данных систем является предельное обобщение видов СЭ и, как следствие, малое число этих видов. Также такие системы способны надежно выявлять лишь небольшое число видов отношений между СЭ. Поэтому для задач, где требуется подробная классификация СЭ с развитой системой логических связей между ними (например, для форм, в которых необходимо идентифицировать разные виды полей) подобный подход не применим.

Отдельно можно выделить системы разбора документов, функционирующие вне систем оптического распознавания. Такие системы в качестве входных данных получают не графические образы страниц, а страницы форматированного, но логически неструктурированного текста – такие данные могут поставляться, например, в файлах формата PDF [13, 14]. Алгоритмы, используемые для разбора таких данных, во многом сходны с алгоритмами анализа документов в системах оптического распознавания. Отличие их состоит в том, что для разбора данных, не пришедших от распознающих модулей, не требуется устойчивости к ошибкам первичных детекторов.

Создание описания модели представления документа определенного типа может быть как полностью ручным [3, 4, 15], так и в разной степени автоматизированным. Подход, частично автоматизирующий создание описания документа на основании одного примера, предложен в [5].

Разработан также подход, использующий автоматическое построение модели документа на основе множества примеров. Надежность данного подхода напрямую зависит от числа примеров, использованных при создании модели. Чем менее жесткой становится структура документа, тем сложнее автоматизировать генерацию его модели представления, и тем больше преимуществ предоставляет задание такой модели вручную [16].

Одно из важных преимуществ ручного задания модели представления – возможность настройки системы распознавания в отсутствие примеров распознаваемого документа.

В настоящей статье представлен подход, основанный на модели представления документа, включающей описание свойств (атрибутов) СЭ и отношений между ними. Данное описание, вместе с растровым изображением документа, подается на вход алгоритма разбора документа. Описание представлено на некоем формальном языке и включает в себя:
– перечисление СЭ документа;
– атрибуты внешнего представления СЭ;
– относительное расположение СЭ на листе.

Результатом работы алгоритма является однозначное сопоставление каждого СЭ с конкретными геометрическими координатами на листе бумаги.

1. Атрибуты структурных элементов

В качестве СЭ документа рассматриваются ключевые слова, общие для всех документов данного вида (как правило, это название самого документа, а также наименования полей), поля, имеющие сходный синтаксис, и линии разграфки, представляющие собой отрезки прямых, параллельные координатным осям.

Среди атрибутов текстовых СЭ, т.е. ключевых слов и полей, можно выделить:
– шрифт, которым напечатан данный элемент;
– кегль (размер шрифта);
– межсимвольный интервал;
– интерлиньяж (для многострочных элементов);
– синтаксис, которому отвечает текст, составляющий данный элемент.

Атрибутами линии разграфки являются:
– ориентация (горизонтальная, вертикальная);
– толщина (жирность);
– длина линии.

Кроме того, в рассматриваемом подходе к атрибутам также отнесена информация о расположении СЭ на странице – расстояния от каждой из границ СЭ границ страницы. По форме данная информация напоминает отношение (о которых речь пойдет ниже), но в силу того, что проверка данных свойств требует анализа лишь одного СЭ, эти сведения отнесены к атрибутам. Предполагается, что размер страницы, соответствующей входному образу, задан изначально – например, соответствует формату А4. Все расстояния, встречающиеся в модели представления документа, даются с учетом этого размера.

Для распознавания всех перечисленных атрибутов предполагается применять соответствующие первичные детекторы.

Не все из вышеуказанных атрибутов использовались при реализации действующей системы разбора документа. Об особенностях существующей реализации речь пойдет ниже.

2. Синтаксис текстовых структурных элементов

Задание такого атрибута как синтаксис текстового СЭ требует отдельного разбора.

Широко применяемым инструментом для задания синтаксисов является язык регулярных выражений (regular expressions). Расширения данного языка, реализованные в существующих системах [17, 18], могли бы покрыть подавляющее большинство задач, возникающих при поиске текстовых СЭ документа. Однако логика расширенных версий этого языка довольно сложна. Реализация также значительно усложняется тем фактом, что на входе модуля сопоставления будут не однозначные текстовые строки, а многоальтернативные строки с весами, полученные от модуля текстового распознавания. Таким образом от модуля сопоставления требуется не двузначный ответ (совпадает/не совпадает), а число, представляющее собой оценку качества совпадения. Это делает реализацию логики сопоставления трудоемкой, а итоговый алгоритм сопоставления – низкоэффективным по скорости.

Поэтому для нашей задачи предлагается упрощенный язык, основанный на операциях конкатенации (обозначаемой знаком &) и альтернативы (которая будет обозначаться знаком |). Элементарной единицей синтаксиса является подстрока. Для уточнения приоритетов используем скобки.

Пример выражения на нашем языке:

«Платежное» & («поручение» | «требование»)

что сопоставляется с образцами «Платежное поручение» и «Платежное требование».

Удобно также оснастить язык многовариантностью на уровне символов, например:

«Номер документа» & [A-Z][0-9][0-9]

что означает, что номер документа состоит из трех символов: ведущей буквы и двух цифр.

Формально такое введение не повысит мощность языка, однако упростит как описание синтаксисов, так и алгоритм сличения с образцами.

Данный упрощенный язык описания синтаксиса строки хорошо подходит для сличения с многоальтернативными строками с весами. При этом для поиска отдельной подстроки используется специальная процедура поиска подстрок в альтернативах распознавания. Эта процедура допускает нежесткое совпадение строк – замену, пропуск букв, лишние буквы, и выдает оценку совпадения в виде числа. Эта же процедура обеспечивает устойчивость подсистемы поиска синтаксисов к опечаткам и ошибкам распознавания.

При поиске синтаксиса, представляющего собой конкатенацию двух синтаксисов, за оценку соответствия строки искомому синтаксису принимается среднее арифметическое оценок каждого операнда с весами, соответствующими длинам строк операндов (такой способ объединения оценок был выбран эмпирически). При поиске синтаксиса, представляющего собой альтернативу двух синтаксисов, за оценку соответствия строки искомому синтаксису принимается наибольшая из оценок соответствия альтернативным синтаксисам. Этот набор правил позволяет написать простую и быструю процедуру сличения синтаксиса с результатом распознавания строки.

С другой стороны, упрощение языка описания синтаксиса ведет к сужению множества строк, которые возможно описать одним синтаксисом. Например, повторное использование ранее найденных групп символов, возможное в расширениях языка регулярных выражений, в предложенном языке невозможно (соответственно, нельзя выявлять повторяющиеся подстроки). Однако для разбора большей части текстовой информации, встречающейся в реальных документах, предложенного языка вполне достаточно.

3. Отношения между структурными элементами

Под отношениями между СЭ понимается их взаимное расположение.

Введем следующие группы отношений.

1. «*Правее*», «*левее*», «*выше*», «*ниже*». Для каждого из данных отношений задаются также пределы расстояний. Например:

   A правее B, минимальное_расстояние=2 мм, максимальное_расстояние=10 мм

   Это означает, что левая граница элемента *A* расположена правее правой границы элемента *B* и расстояние между этими границами находится в пределах от 2 до 10 мм.
2. «*Справа*», «*слева*», «*над*», «*под*». Данная группа отношений отличается от предыдущей, тем, что указывает также на то, что элементы близки (перекрываются) либо по горизонтальной коор-

динате (для отношений «*над*» и «*под*»), либо по вертикальной (для отношений «справа» и «слева»). Для данных отношений так же, как и для предыдущей группы, задаются пределы расстояний.

3. Только для линий разграфки вводится отношение «*соединены*». Соединенными считаются линии, у одной из которых один из концов совпадает с одним из концов другой. Т-стыки рассматриваются как соединение трех линий в одной точке.

Также, для удобства использования введем составные отношения «*до*» и «*после*». «*До*» означает «*слева*» либо «*выше*». Аналогично, «*после*» – это «*справа*» либо «*ниже*».

4. Граф отношений

Введенные отношения обладают симметрией с точки зрения приоритета входящих в них СЭ. Т.е. если элементы A и B связаны отношением, то при известном расположении A отношение наложит ограничения на расположение B, но аналогично верно и обратное: при известном расположении B ограничения будут наложены на A.

Однако в рассматриваемой модели отношения будут считаться асимметричными: каждому из двух участвующих в отношении СЭ будет назначена роль ведущего либо роль зависимого. Т.е. если утверждается «*A правее B*», то в дальнейшем будет рассматриваться, какие ограничения на A накладывает расположение B, но не наоборот.

Таким образом, граф отношений является ориентированным графом. Направления ребер будут полагаться от зависимых СЭ к ведущим.

На граф также накладывается дополнительное ограничение: граф не должен содержать циклы. Это ограничение существенно для сходимости нижеприведенного алгоритма.

Из вышесказанного следует, что построение графа отношений, описывающего конкретный тип документа, не является однозначным. Один и тот же тип документа может описываться множеством различных графов. Принципы, которыми следует руководствоваться при создании графа, вытекают из алгоритма разбора документа и будут описаны ниже.

5. Алгоритм разбора документа

Итак, имеется модель документа, состоящая из СЭ с заданными атрибутами и отношениями.

Общая структура алгоритма разбора документа на основе этой информации:

1. Выбрать произвольный СЭ, для которого поиск еще не выполнялся.
2. Если таких СЭ не осталось, завершить алгоритм.
3. Осуществить поиск выбранного СЭ, получить множество кандидатов с весами.
4. В качестве окончательного результата поиска СЭ выбрать кандидат с наибольшим весом.
5. Перейти к шагу 1.

Порядок выбора элементов на первом шаге несущественен, т.к., как будет описано ниже, алгоритм рекурсивно вызывает поиск для всех ведущих СЭ, которые еще не были найдены. Таким образом, реальный порядок поиска СЭ всегда будет от ведущих к зависимым. Различия в порядке поиска могут быть только для СЭ, находящихся в разных компонентах связности графа, т.е. полностью независимых друг от друга.

Алгоритм поиска СЭ, вызываемый на шаге 3, состоит из следующих шагов:

1. Поиск СЭ по атрибутам.
2. Поиск СЭ по отношениям.

Поиск по атрибутам генерирует начальное множество альтернативных местонахождений СЭ с оценками. Поиск по отношениям уточняет оценку каждой альтернативы в зависимости от степени удовлетворения отношениям, в которых СЭ участвует.

После завершения для каждого СЭ обоих этапов поиска выбираются наилучшие кандидаты для всех СЭ.

6. Поиск структурного элемента по атрибутам

На данном этапе производится поиск СЭ вне зависимости от его отношений с другими СЭ. Входные данные этого этапа – набор текстовых строк, распознанных на странице (в многовариантном формате с весами), а также набор линий, найденных на странице. Результатом является набор кандидатов на расположение данного СЭ (список координат прямоугольников возможного его расположения).

Ключевым атрибутом для поиска текстового СЭ является его синтаксис. Первым этапом поиска элемента по атрибутам является поиск строк, отвечающих заданному синтаксису. Этот поиск осуществляется в два этапа: поиск всех подстрок, входящих в синтаксис, и составление всевозможных цепочек подстрок, отвечающих синтаксису. При поиске подстрок учитываются все альтернативы распознавания с весами. В результате найденные подстроки также обладают весами. При склеивании подстрок в цепочки веса интегрируются. В

результате получаем первичный набор кандидатов для данного СЭ.

Для линии разграфки первичными кандидатами являются все линии данной ориентации (горизонтальной или вертикальной).

Далее, для каждого кандидата из первичного набора проверяем его соответствие прочим атрибутам: кеглю и интерлиньяжу для текстов, длине и толщине – для линий. По результатам проверки корректируем веса элементов. Кандидаты, не прошедшие некоторый порог веса – исключаем из множества.

Итоговое множество кандидатов является входными данным для поиска СЭ по отношениям.

7. Поиск структурного элемента по отношениям

Поиск СЭ по отношениям состоит в анализе соответствия каждого из кандидатов набору отношений, для которых данный СЭ является зависимым (ведущие СЭ этих отношений назовем ведущими СЭ для данного элемента).

Просматриваются все СЭ, ведущие для данного. Для тех из них, для которых еще не осуществлялся поиск – выполняем поиск. Таким образом, мы анализируем отношения только с уже найденными СЭ. Запуск поиска для ведущего СЭ является рекурсивным (т.к. мы в данный момент уже находимся внутри алгоритма поиска СЭ). Чтобы данная рекурсия была сходящейся, и было введено требование ацикличности графа отношений.

Далее, для каждого ведущего СЭ берем набор его кандидатов. Каждый из данных кандидатов, вследствие наличия отношения с искомым СЭ, налагает некоторые ограничения на расположение искомого СЭ. Данные ограничения являются взаимоисключающими, т.к. соответствуют взаимоисключающим кандидатам ведущего СЭ. Назовем множество взаимоисключающих ограничений, полученных через одно отношение с ведущим СЭ, набором альтернативных ограничений (а его элементы, соответственно, альтернативными ограничениями).

Каждое из альтернативных ограничений имеет некий вес – соответствующий весу того кандидата ведущего СЭ, для которого сгенерировано данное ограничение.

Получаем для искомого СЭ множество наборов альтернативных ограничений (по одному набору на каждое отношение с каждым ведущим СЭ).

Теперь пройдем по множеству кандидатов искомого СЭ, полученных в поиске по атрибутам. Для каждого из кандидатов оценим его соответствие всему множеству наборов альтернативных ограничений. Для этого в каждом наборе альтернативных ограничений найдем альтернативное ограничение наибольшего веса, которому наш кандидат отвечает. Сумма весов таких лучших альтернатив, деленная на максимально возможный вес при данном числе ведущих СЭ – и есть итоговая оценка кандидата искомого СЭ.

В результате получаем множество кандидатов СЭ с весами. Кандидаты с наибольшим весом для каждого СЭ и являются результатом работы алгоритма разбора документа.

8. Принципы построения графа отношений

Граф отношений строится вручную и его построение не является однозначным, однако от выбора графа зависит итоговая надежность разбора документа. Из вышеприведенного алгоритма можно вывести следующие принципы построения графа, обеспечивающие максимальную устойчивость алгоритма.

Прежде всего, следует заметить, что для разных СЭ требуется разная надежность поиска, а именно, в подавляющем большинстве случаев пользователю системы требуется только надежное выделение полей, а ошибки в выделении ключевых слов и линий разграфки несущественны.

С другой стороны, чем больше ведущих СЭ существует для данного СЭ, тем надежнее поиск данного СЭ, т.к. опирается на больший объем информации (в данном случае мы рассматриваем транзитивно замкнутые отношения ведущий-зависимый, т.е. если A зависит от B а B зависит от C, то A зависит от C).

Следовательно, СЭ, наиболее важные для детекции – поля – следует располагать насколько возможно близко к вершинам-истокам ориентированного графа отношений.

Также можно обратить внимание на то, что если рассматривать СЭ вне отношений, т.е. как сущность, располагающую только атрибутами, то надежность поиска каждого СЭ будет зависеть от его атрибутов. Например, надежнее будут выявляться более длинные текстовые СЭ с более строгим синтаксисом. Т.е. устойчивее выделяются те СЭ, при поиске которых сопоставляется больше информации. В общем случае граф отношений следует проектировать так, чтобы более надежно выявляемые СЭ были ведущими для менее надежно выявляемых, т.к. это увеличивает средний объем информации, используемой для поиска каждого СЭ, т.е. повышает общую устойчивость разбора документа.

9. Особенности реализации

Система создана на основе существующей системы распознавания текстовых страниц CuneiForm. Выходом системы распознавания страниц является неструктурированное множество найденных на листе строк, для каждой из которых доступно множество вариантов ее распознавания с оценками. Система распознавания является шрифтонезависимой и не сообщает какой-либо информации о шрифте, которым напечатан текст. В связи с этим информация о типе шрифта изначально не использовалась в качестве атрибута при разборе. Добавление такой возможности было признано нецелесообразным с точки зрения соотношения затрат и прогнозируемого улучшения качества разбора. Действительно, в стандартах большинства реальных документов не указан требуемый шрифт и разные организации используют в документах разные шрифты. Кроме того, в рамках одного документа в подавляющем большинстве случаев используется лишь один шрифт. Поэтому тип шрифта не представляет собой информации, существенной для различения СЭ документа.

Сходная ситуация сложилась с учетом начертания шрифта (жирный, курсив) – начертание одних и тех же текстовых СЭ одного и того же типа документа менялось от экземпляра к экземпляру документа, поэтому реализовывать распознавание этих атрибутов сочли нецелесообразным.

Информация же о кегле полезна для детекции частей документа, однако ее легко получить из размеров охватывающих прямоугольников распознанных символов, без какого-либо дополнительного анализа на этапе распознавания.

Вместо атрибута, задающего межсимвольный интервал, был введен атрибут, указывающий пределы ширины строки по отношению к кеглю символа. Причина такой замены в том, что по распечатке документа, в отсутствие исходного макета, межсимвольный интервал определить непросто, в то время как ширину текста измерить легко.

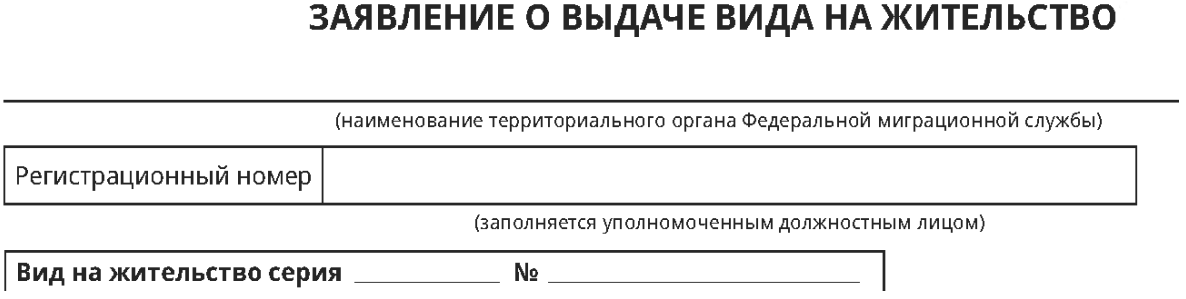

Рис.1. Фрагмент документа, описанный на языке атрибутов и отношений

10. Пример описания документа на языке атрибутов и отношений

Одно из практических применений реализованной системы – разбор документа в рамках системы автоматизированного ввода заявлений на выдачу документов, удостоверяющих личность.

На рис.1. приведен фрагмент заявления о выдаче вида на жительство. Ниже приведено описание этого фрагмента на языке атрибутов и отношений. Так выглядит описание присутствующих на изображении линий разграфки:

ЛИНИЯ линия1 ГОРИЗ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МИН = 0.7 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.2;
ЛИНИЯ линия2_1 ГОРИЗ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МИН = 0.6 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.25;
ЛИНИЯ линия2_2 ГОРИЗ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МИН = 0.2 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.25;
ЛИНИЯ линия2_3 ГОРИЗ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МИН = 0.6 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.3;
ЛИНИЯ линия2_4 ГОРИЗ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МИН = 0.2 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.3;
ЛИНИЯ линия2_5 ВЕРТ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МАКС = 0.1
ЛИНИЯ линия2_6 ВЕРТ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МАКС = 0.1
ЛИНИЯ линия2_7 ВЕРТ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МАКС = 0.1

ЛИНИЯ линия3_1 ГОРИЗ ТОЛЩИНА_МАКС = 2 ДЛИНА_ОТН_МИН = 0.4 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.35;
ЛИНИЯ линия3_2 ВЕРТ ТОЛЩИНА_МАКС = 2 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.35 ОТСТУП_ЛЕВ_ОТН_МИН = 0.2;
ЛИНИЯ линия3_3 ВЕРТ ТОЛЩИНА_МАКС = 2 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.35 ОТСТУП_ЛЕВ_ОТН_МИН = 0.3;
ЛИНИЯ линия4 ГОРИЗ ТОЛЩИНА_МАКС = 1 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.4;
ЛИНИЯ линия5 ГОРИЗ ТОЛЩИНА_МАКС = 1 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.4;

Для каждой линии задана ориентация («*ГОРИЗ*» или «*ВЕРТ*»), а также ограничения на толщину и длину линий. Для части линий также заданы пределы отступа от тех или иных границ изображения. Суффикс «*_ОТН_*», присутствующий в названии некоторых атрибутов указывает на то, что величина указана в долях от ширины либо высоты листа.

Далее следует описание меток (статических текстов) формы, с заданием синтаксиса и ограничениями на кегль:

МЕТКА заголовок СИНТАКСИС = “ЗАЯВЛЕНИЕ О ВЫДАЧЕ ВИДА НА ЖИТЕЛЬСТВО“ КЕГЛЬ_МИН = 12 ОТСТУП_ВЕРХ_ОТН_МАКС = 0.2;
МЕТКА наименование_терр СИНТАКСИС = “(наименование территориального органа“ & (“Федеральной миграционной службы)“ | “ФМС)”) КЕГЛЬ_МАКС = 10;
МЕТКА регистрационный СИНТАКСИС = “Регистрационный номер” КЕГЛЬ_МИН = 10 КЕГЛЬ_МАКС = 12;
МЕТКА заполняется СИНТАКСИС = “(заполняется уполномоченным должностным лицом)“ КЕГЛЬ_МАКС = 10;
МЕТКА внж_серия СИНТАКСИС = “Вид на жительство серия“ КЕГЛЬ_МИН = 10 КЕГЛЬ_МАКС = 12;
МЕТКА внж_номер СИНТАКСИС = “№“ | “номер“ КЕГЛЬ_МИН = 10 КЕГЛЬ_МАКС = 12;

Далее идет описание полей, подлежащих распознаванию. С точки зрения системы разбора документа поля ничем не отличаются от статических текстов. Различие введено для удобства обработки результатов разбора модулями финального распознавания.

ПОЛЕ поле_терр_орган;
ПОЛЕ поле_рег_номер;
ПОЛЕ поле_внж_серия СИНТАКСИС=[0-9][0-9];
ПОЛЕ поле_внж_номер СИНТАКСИС=[0-9][0-9][0-9][0-9][0-9][0-9][0-9];

Для полей, представляющих серию и номер вида на жительство, заданы синтаксисы, в соответствии с которыми серия состоит из двух произвольных цифр, а номер – из семи произвольных цифр.

Далее описаны отношения между СЭ формы:

ОТНОШЕНИЕ линия1 ПОД заголовок РАССТ_ОТН_МАКС = 0.1;
ОТНОШЕНИЕ наименование_терр ПОД линия1 РАССТ_ОТН_МАКС = 0.01;
ОТНОШЕНИЕ линия2 ПОД наименование_терр РАССТ_ОТН_МАКС = 0.1;
ОТНОШЕНИЕ регистрационный ПОД линия2_1 РАССТ_ОТН_МАКС = 0.01;
ОТНОШЕНИЕ регистрационный СПРАВА линия2_5;
ОТНОШЕНИЕ регистрационный СЛЕВА линия2_6;
ОТНОШЕНИЕ регистрационный НАД линия2_4 РАССТ_ОТН_МАКС = 0.01;
ОТНОШЕНИЕ заполняется ПОД линия2_3 РАССТ_ОТН_МАКС = 0.01;
ОТНОШЕНИЕ линия3_1 НИЖЕ заполняется;
ОТНОШЕНИЕ внж_серия ПОД линия3_1;
ОТНОШЕНИЕ линия4 ПОД линия3_1;
ОТНОШЕНИЕ линия5 СПРАВА линия4;

ОТНОШЕНИЕ внж_серия ВЫШЕ линия5;
ОТНОШЕНИЕ внж_номер СПРАВА внж_серия;
ОТНОШЕНИЕ линия2_2 СОЕД линия 2_1;
ОТНОШЕНИЕ линия2_3 СОЕД линия 2_4;
ОТНОШЕНИЕ линия2_5 СОЕД_ВЕРХЛЕВО линия 2_1;
ОТНОШЕНИЕ линия2_6 СОЕД_ВЕРХПРАВО линия 2_1;
ОТНОШЕНИЕ линия2_6 СОЕД_ВЕРХЛЕВО линия 2_2;
ОТНОШЕНИЕ линия2_7 СОЕД_ВЕРХПРАВО линия 2_2;
ОТНОШЕНИЕ линия2_5 СОЕД_НИЗЛЕВО линия 2_3;
ОТНОШЕНИЕ линия2_6 СОЕД_НИЗПРАВО линия 2_3;
ОТНОШЕНИЕ линия2_6 СОЕД_НИЗЛЕВО линия 2_4;
ОТНОШЕНИЕ линия2_7 СОЕД_НИЗПРАВО линия 2_4;
ОТНОШЕНИЕ поле_терр_орган ПОД заголовок;
ОТНОШЕНИЕ поле_терр_орган НАД линия1;
ОТНОШЕНИЕ поле_рег_номер СПРАВА линия2_6;
ОТНОШЕНИЕ поле_внж_серия СПРАВА внж_серия;
ОТНОШЕНИЕ поле_внж_серия СЛЕВА внж_номер;
ОТНОШЕНИЕ поле_внж_серия НАД линия4;
ОТНОШЕНИЕ поле_внж_номер СПРАВА внж_номер;
ОТНОШЕНИЕ поле_внж_номер НАД линия5;
ОТНОШЕНИЕ поле_внж_номер СЛЕВА линия3_3;

Указаны геометрические отношения между СЭ. О различиях между отношениями «*ПОД*» и «*НИЖЕ*», а также «*НАД*» и «*ВЫШЕ*» было сказано ранее. Для некоторых отношений также заданы пределы расстояния между СЭ (имеется в виду расстояние по той координате, которая соответствует направлению отношения, т.е., например, для отношения «*ПОД*» берется расстояние по у-координате).

Для линий также заданы отношения соединения «*СОЕД*». Для двух линий разграфки одной ориентации дополнительных уточнений не требуется – соединение таких линий означает, что линии соприкасаются концами (линии одинаковой толщины при этом будут фактически сливаться в одну). Для линий разной ориентации указано, какими именно концами они соприкасаются (например, «*СОЕД_НИЗЛЕВО*» означает, что горизонтальная линия касается левым концом нижнего конца вертикальной линии).

Порядок описания СЭ и отношений не важен, однако важен порядок следования СЭ в отношениях: Первый упомянутый в отношении СЭ считается зависимым, второй – ведущим. Как было сказано выше, отношения должны образовывать граф без циклов.

Если положить ориентацию ребер графа от зависимого СЭ к ведущему, то поиск СЭ по атрибутам начнется с любой вершины-стока графа.

Данный метод разбора документа был реализован и применялся в рамках системы автоматизированного ввода для документов, разбор которых способами [1, 2, 5], использовавшимися ранее, оказался невозможен. С другой стороны, метод давал более устойчивые результаты и работал быстрее, чем медоды [3, 4]. Предложенная система успешно эксплуатировалась для распознавания более 20 типов документов, среди которых можно назвать Счета-фактуры, товарные накладные, акты приемки-передачи, заявления на выдачу документов, удостоверяющих личность и другие.

Заключение

В настоящей работе предложена система разбора документов в рамках системы автоматизированного ввода с использованием технологий компьютерного зрения. Особенностью системы является ее способность работать с документами, структурные элементы которых не имеют фиксированного расположения на бумаге, а также с документами, примеры которых не представлены. Система показала эффективность и практическую применимость для автоматизированного ввода документов. Также система показала высокую скорость работы, хорошую устойчивость результатов разбора, а для ряда документов при ее применении устойчивость оказалась значительно выше, чем в ранее существовавших системах.

За счет средств гранта РНФ (проект #14-50-00150) предложен подход распознавания, основанный на описании объекта через свойства его элементов, а также предложен алгоритм разбора документа с целью распознавания, реализующий данный подход.

Система основана на алгоритме, предполагающем описание документа набором структурных элементов с заданными атрибутами и отношениями между структурными элементами. Благодаря введенному требованию ацикличности графа отношений, разбор свелся к последовательному поиску структурных элементов с учетом расположения ранее найденных элементов. Таким образом, скорость работы алгоритма является линейной относительно числа структурных элементов. Быстрота данного алгоритма является одним из важнейших его преимуществ, в сравнении с существующими методами разбора документов, структурные элементы которых не имеют фиксированного расположения на бумаге.

Результат разбора данным алгоритмом является зависимым от построения графа отношений, который задается вручную. Для устранения данного недостатка в дальнейшем предполагается автоматизировать построение оптимального ориентированного ациклического графа отношений на основе заданного вручную неориентированного графа отношений. Этой задаче планируется посвятить будущие исследования.

Литература

1. *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.
2. *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.
3. *Постников В.В., Марченко А.Е., Шоломов Д.Л.*, Разбор структурированного документа в модели с нечеткой логикой. // В сб. «Документооборот. Концепции и инструментарий», Москва, М.: URSS, 2004.
4. *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.
5. *Постников В.В.* Автоматическая идентификация и распознавание структурированных документов : диссертация ... кандидата технических наук : 05.13.01 Москва, 2001 126 с. : 61 02-5/365-8
6. *Eugene Borovikov*, “A survey of modern optical character recognition techniques” arXiv preprint arXiv:1412.4183, 2014.
7. *Olivier Augereau, Nicholas Journet, Jean-Philippe Domenger,* “Semi-structured document image matching and recognition”, Proc. SPIE 8658, Document Recognition and Retrieval XX, 865804 (4 February 2013).
8. *Bertrand Coüasnon, Aurélie Lemaitre,* “Recognition of Tables and Forms”, Handbook of Document Image Processing and Recognition, 2014.
9. *Cattoni R., Coianiz T., Messelodi S., Modena C.M.,* “Geometric Layout Analysis Techniques for Document Image Understanding: a Review”, Technical Report, IRST, Trento, Italy, 1998.
10. *Thomas M Breuel,* “High performance document layout analysis”, Proceedings of the Symposium on Document Image Understanding Technology, 2003.
11. *Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, C. Lee Giles.* “Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017).
12. *Tatsuhiko Kagehiro, Hiromichi Fujisawa,* “Multiple Hypotheses Document Analysis”, Machine Learning in Document Analysis and Recognition, 2008.
13. *Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, Gully APC Burns,* “Layout-aware text extraction from full-text PDF of scientific articles”, Source Code for Biology and Medicine 7(1), 2012.
14. *Hui Chao, Jian Fan,* “Layout and Content Extraction for PDF Documents. 2004. Layout and content extraction for pdf documents”, International Workshop on Document Analysis Systems. Springer, 2004.
15. *Niyogi D. and Srihari S.N.,* “Knowledge-based derivation of document logical structure”, Proceedings of the 3rd International Conference on Document Analysis and Recognition – ICDAR, *1995.*
16. *Голубев С.В.*, Распознавание структурированных документов на основе машинного обучения. // Бизнес-информатика. – № 2 (16), 2011.
17. *“Regular Expressions”*. The Single UNIX ® Specification, Version 2. [electronic resource] // The Open Group [official website]. URL: http://pubs.opengroup.org/onlinepubs/007908799/xbd/re.html (accessed: 1.09.2017)

18. *Perl-compatible Regular Expressions* (revised API: PCRE2) [electronic resource] // PCRE - Perl Compatible Regular Expressions [official website]. URL: http://pcre.org/current/doc/html/ (accessed: 1.09.2017)

Марченко Алексей Евгеньевич. Разработчик программного обеспечения ООО «Когнитивные технологии». Окончил в 2002 г. МФТИ (ГУ). Количество печатных работ: 6. Область научных интересов: информационные технологии, зрительные системы, распознавание образов. E-mail: alexey@cognitive.ru

Ершов Егор Иванович. М.н.с. ИППИ РАН им. А.А. Харкевича. Окончил в 2014 г. МФТИ (ГУ). Количество печатных работ: 21. Область научных интересов: обработка изображений, компьютерное зрение, разработка и исследование алгоритмов распознавания образов. E-mail: ershov@iitp.ru

Гладилин Сергей Александрович. С.н.с. ИППИ РАН им. А.А. Харкевича. К.т.н. Окончил в 2002 г. МГУ им. М.В. Ломоносова. Количество печатных работ: 24. Область научных интересов: зрительные системы, обработка изображений, распознавание образов. Email: gladilin@iitp.ru

System of parsing of documents specified by structure item attributes and relations between the items

A.E. Marchenko, E.I. Ershov, S.A. Gladilin

Abstract. Within the problem of document recognition with computer vision technologies the problem of finding the correspondence between the structure items of a document and their printed images that have no strict locations is concerned. An approach based on document description with attributes of its structure items and relations between the items is proposed. An algorithm of document parsing using this approach is proposed. A system implementing document parsing based on this approach is described.

Keywords: document parsing, structure item, relations between items, item attributes, parsing algorithm.

References

1. *Usilin S.A., Nikolaev D.P. and Postnikov V.V.* 2008. Bystryy algoritm sovmeshcheniya izobrazheniy dokumentov v proizvol'noy geometricheskoy modeli [A fast algorithm of document image superposition in an arbitrary geometrical model]. Trudy konferentsii "Informatsionnye tehnologii i sistemy [Conference "Information Technologies and Systems" Preceedings]. Gelendzhik. 471 – 477.
2. *Bezmaternyh P.V., Nikolaev D.P. and Postnikov V.V*. 2008. Metod identifikatsii tipa documenta po strukture ego proektsiy na koordinatnye osi [Method of identifying of type of a document by structure of its projections to coordinate axes]. Trudy konferentsii "Informatsionnye tehnologii i sistemy [Conference "Information Technologies and Systems" Preceedings]. Gelendzhik. 498 – 501.
3. *Postnikov V.V., Marchenko A.E. and Sholomov D.L.* 2004. Razbor strukturirovannogo dokumenta v modeli s nechetkoy logikoy [Structured document parsing in a model with fuzzy logic]. Dokumentooborot. Kontseptsii i instrumentariy [Document Flow: Concepts and Toolkits].
4. *Postnikov V.V.* 2001. Avtomaticheskaya identifikatsiya i raspoznavanie strukturirovannyh dokumentov [Automatic structured documents identification and recognition]. C. Sc. Diss. Moscow. 126 p.
5. *Postnikov V.V.* and Marchenko A. E. 2005. CFML: yazyk opisaniya mnogostranichnyh strukturirovannyh dokumentov dlya ih identifikatsii i raspoznavaniya [CFML: a language of description of multipage structured documents for their identification and recognition]. Matematicheskie metody raspoznavaniya obrazov (MMRO-12): Sbornik dokladov 12-y Vserossiyskoy Konferentsii [Mathematical Methods of Pattern Recognition: The 12th All-Russian Conference Preceedings].
6. *Eugene Borovikov*, "A survey of modern optical character recognition techniques" arXiv preprint arXiv:1412.4183, 2014.
7. *Olivier Augereau, Nicholas Journet, Jean-Philippe Domenger*, "Semi-structured document image matching and recognition", Proc. SPIE 8658, Document Recognition and Retrieval XX, 865804 (4 February 2013).
8. *Bertrand Coüasnon, Aurélie Lemaitre*, "Recognition of Tables and Forms", Handbook of Document Image Processing and Recognition, 2014.
9. *Cattoni R., Coianiz T., Messelodi S., Modena C.M.,* "Geometric Layout Analysis Techniques

for Document Image Understanding: a Review", Technical Report, IRST, Trento, Italy, 1998.
10. *Thomas M Breuel*, "High performance document layout analysis", Proceedings of the Symposium on Document Image Understanding Technology, 2003.
11. *Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, C. Lee Giles*. "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017).
12. *Tatsuhiko Kagehiro, Hiromichi Fujisawa*, "Multiple Hypotheses Document Analysis", Machine Learning in Document Analysis and Recognition, 2008.
13. *Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy*, Gully APC Burns, "Layout-aware text extraction from full-text PDF of scientific articles", Source Code for Biology and Medicine 7(1), 2012.
14. *Hui Chao, Jian Fan*, "Layout and Content Extraction for PDF Documents. 2004. Layout and content extraction for pdf documents", International Workshop on Document Analysis Systems. Springer, 2004.
15. *Niyogi D. and Srihari S.N.*, "Knowledge-based derivation of document logical structure", Proceedings of the 3rd International Conference on Document Analysis and Recognition – ICDAR, 1995.
16. *Golubev S.V,* Raspoznavanie strukturirovannyh dokumentov na osnove mashinnogo obucheniya [Recognition of Structured Documents Based on Machine Learning]. Biznes-informatika [Business Informatics]. – № 2 (16), 2011.
17. *"Regular Expressions"*. The Single UNIX ® Specification, Version 2. [electronic resource] // The Open Group [official website]. URL: http://pubs.opengroup.org/onlinepubs/007908799/xbd/re.html (accessed: 1.09.2017)
18. *Perl-compatible Regular Expressions* (revised API: PCRE2) [electronic resource] // PCRE - Perl Compatible Regular Expressions [official website]. URL: http://pcre.org/current/doc/html/ (accessed: 1.09.2017)

Marchenko Alexey Evgenievich. Software developer LLC "Cognitive Technologies." He graduated in 2002 from MIPT (SU). Number of publications: 6. Area of scientific interests: information technology, visual systems, pattern recognition. E-mail: alexey@cognitive.ru

Ershov Egor Ivanovich. Junior research associate. IITP RAS named after. A. A. Kharkevich. Graduated in 2014 MIPT (SU). Number of publications: 21. Research interests: image processing, computer vision, development and study of image recognition algorithms. E-mail: ershov@iitp.ru

Gladilin Sergey Alexandrovich. Higher senior officer. IITP RAS named after. A. A. Kharkevich. Ph.D. He graduated in 2002 from Moscow State University named after M.V. Lomonosov. Number of publications: 24. Research interests: visual systems, image processing, image recognition. E-mail: gladilin@iitp.ru