Особенности контекстного распознавания российского заграничного паспорта в системе Cognitive Passport

Д. Г. Слугин

Институт системного анализа Российской академии наук,
Россия, 117312 Москва, пр. 60-летия Октября, 9

Статья посвящена задаче контекстной обработки российского заграничного паспорта в системе Cognitive Passport. Рассмотрены особенности распознавания, предложен алгоритм геометрического разделения полей, а также дополнительный контроль и восстановление значений при помощи машиночитаемой зоны.

Введение

Задача распознавания российского заграничного паспорта (далее — загранпаспорта) имеет большое практическое значение, как в органах государственной власти, так и в частных организациях. По данным Федеральной миграционной службы, около 80 % граждан России имеют загранпаспорт, следовательно, речь идет о десятках миллионов документов. Сфера применения загранпаспорта не ограничивается только удостоверением личности при пересечении границы, он может быть использован в качестве дополнительного документа при взятии кредита или осуществлении платежей через банки или другие финансовые учреждения, а также как средство идентификации личности гражданина РФ. Таким образом, область использования систем распознавания загранпаспортов достаточно широка: контрольно-пропускные пункты на Государственной границе, банки и кредитные организации, туристические фирмы, регистратуры и проходные пункты предприятий, салоны связи и др. От скорости и качества распознавания загранпаспорта напрямую зависит количество обрабатываемых документов в единицу времени, снижение расходов на персонал, повышение конкурентоспособности предприятий, уменьшение времени ожидания для клиентов.

В системе Cognitive Passport загранпаспорт в процессе своей обработки проходит несколько этапов: снятие цветного фона, удаление помех, бинаризация, определение типа документа, привязка элементов, нахождение полей, распознавание букв и формирование строк, контекстное распознавание. Именно особенностям последнего этапа — контекстного распознавания — и посвящена данная статья. Этот этап характеризуется тем, что к моменту его вызова уже известно геометрическое расположение букв в полях, их возможные альтернативы и вероятности распознавания. Требуется на основе этой информации принять решение, какие данные будут окончательно сформированы в выходных полях документа для проверки их оператором и/или сохранения в информационную систему. В системе Cognitive Passport на этапе контекстного распознавания можно условно выделить следующие основные особенности: распознавание зоны данных в виде общего поля, геометрическое разделение полей, обработка машиночитаемой зоны, корректировка значений полей. Рассмотрим их более подробно.

1. Распознавание зоны данных как общего поля

В загранпаспорте (рис. 1) можно выделить две основные зоны, в которых расположена информация о документе: зона данных (верхний прямоугольник) и машиночитаемая зона (нижний прямоугольник).

Рис. 1. Зоны загранпаспорта

Машиночитаемая зона может отсутствовать, подробнее она будет рассмотрена ниже. Зона данных включает в себя следующие поля: номер документа, фамилию владельца и ее транслитерацию, имя и отчество владельца и транслитерацию имени, пол, дату рождения, место рождения, дату выдачи, дату окончания срока действия, а также орган, выдавший документ. Транслитерация — это способ преобразования слов на русском языке в англоязычные на основе замещения русских букв на их английские аналоги или сочетания. Более подробно с правилами транслитерации можно ознакомиться в [2].

На документе присутствует ряд дополнительных полей, но их значения для российского загранпаспорта фиксированы, и поэтому они не распознаются. Это поля: тип документа (значение «P»), код государства («RUS»), гражданство («Российская Федерация») и личный код (не заполняется). Особенностью российского загранпаспорта является то, что шрифт и расположение впечатываемых данных жестко не определены, поля могут «плавать» в достаточно широких пределах и «наезжать» на статические элементы документа. В результате распознавание каждого поля по отдельности не совсем оправданно, так как приводит к достаточному количеству ошибок наводки и снижению качества распознанных символов. Поэтому было предложено решение — распознавать зону данных как одно большое поле (назовем его общим полем), а затем, используя геометрические характеристики полей, определять их местоположение и информацию в них. Отметим, что номер документа все же распознается отдельно, это связано с тем, что его печать производится типографским способом в точно отведенное место, а также из-за особенностей цвета и шрифта, что позволяет получить высокое качество распознавания с минимальным количеством ошибок. В дальнейшем будем считать, что номер документа не входит в общее поле.

2. Геометрическое разделение полей

Поля, расположенные внутри общего поля, геометрически можно представить в виде следующей схемы (рис. 2).

Для каждого поля можно описать его положение относительно других полей, например дата рождения ниже имени, выше пола и левее места рождения. Рассмотрим алгоритм геометрического разделения полей. Общее поле представляет собой массив букв и их альтернатив с вероятностями распознавания и координатами прямоугольников букв. Для геометрического разделения мы будем использовать только координаты, сами значения букв и их альтернативы будут добавляться в соответствующие поля.

Распознавание общего поля идет слева направо, сверху вниз. Таким образом, последовательно двигаясь по буквам и используя определенный критерий, мы либо прибавляем очередную букву к текущему полю, либо начи-

Рис. 2. Геометрическое расположение полей

наем формировать новое. Критерием может являться расстояние от буквы до текущего поля: если расстояние достаточно велико, то начинаем формирование нового поля, иначе расширяем прямоугольник текущего поля так, чтобы он включал в себя прямоугольник буквы. «Достаточно велико» означает в данном случае — на определенную величину, которая может быть заранее рассчитана эмпирически на наборе паспортов или динамически для каждого конкретного паспорта. В системе Cognitive Passport использован следующий критерий: расстояние по горизонтали — более двух средних размеров символа и по вертикали — более половины среднего размера символа на документе.

Расстояние между прямоугольником P и буквой L по горизонтали или вертикали рассчитывается как минимальное расстояние между сторонами в соответствующем направлении (рис. 3). Расширенный прямоугольник S, включающий в себя P и L, показан пунктиром.

Таким образом, в результате геометрического разделения мы должны получить шесть полей слева и два–три поля справа (в зависимости от особенностей надпечатки поля «Место рождения», в одну строку или в две, с большим интервалом по вертикали или нет). Однако бывают ситуации, когда из-за неточной наводки общего поля, слипшихся строк данных или неправильной печати на паспорте число полей отличается от искомого. В таком случае нужна дополнительная идентификация полей, уже с учетом данных внутри каждого поля. В качестве такой идентификации может выступать длина строки, а также его буквенный состав с альтернативами. Например, поле «Пол» составляет не более трех символов (возможные зна-

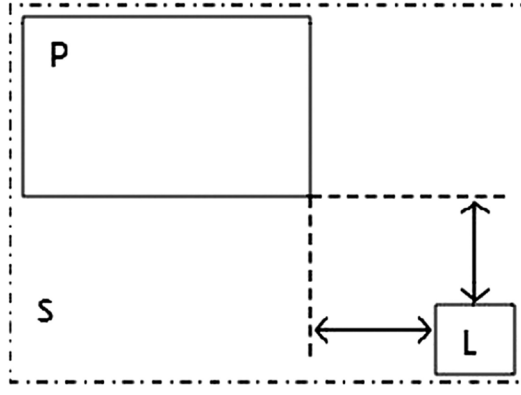

Рис. 3. Расстояние между прямоугольниками

чения M, M/M, M/F), поля дат состоят не более чем из десяти символов и только из чисел и символа точка («.»), поле «Место рождения» часто включает в себя слово «USSR» (для людей, родившихся до 1992 года). Таким образом, можно с большой вероятностью идентифицировать часть полей, а используя геометрические отношения выше/ниже и левее/правее, — и остальные поля.

3. Обработка машиночитаемой зоны

Машиночитаемая зона представляет собой две строки текста и расположена в специально отведенной области загранпаспорта. Каждая строка содержит фиксированное количество символов (44 символа), шрифт печати OCR-B стандарта ISO 1073–2, размер шрифта и набор символов задан. Расположение и размер зоны жестко определены и не допускают серьезных отклонений. Поэтому качество распознавания машиночитаемой зоны очень высокое и приближается к 100 %.

Как видно из структуры, представленной на рис. 4, в ней содержатся следующие значения, соответствующие значениям полей зоны данных: имя и фамилия владельца (транслитерация), пол, дата рождения и дата окончания срока действия. Также указан номер документа. Для части загранпаспортов надпечатка может отсутствовать.

Для проверки корректности распознанных строк присутствуют контрольные цифры, как индивидуальные для некоторых значений, так и общие для всей строки. Более подробно о кодировании строк и контрольных цифрах можно прочитать в [3]. Даты рождения и срок действия документа представлены в формате ГГММДД, однако их можно легко перевести в формат с четырехзначным годом по следующему правилу:

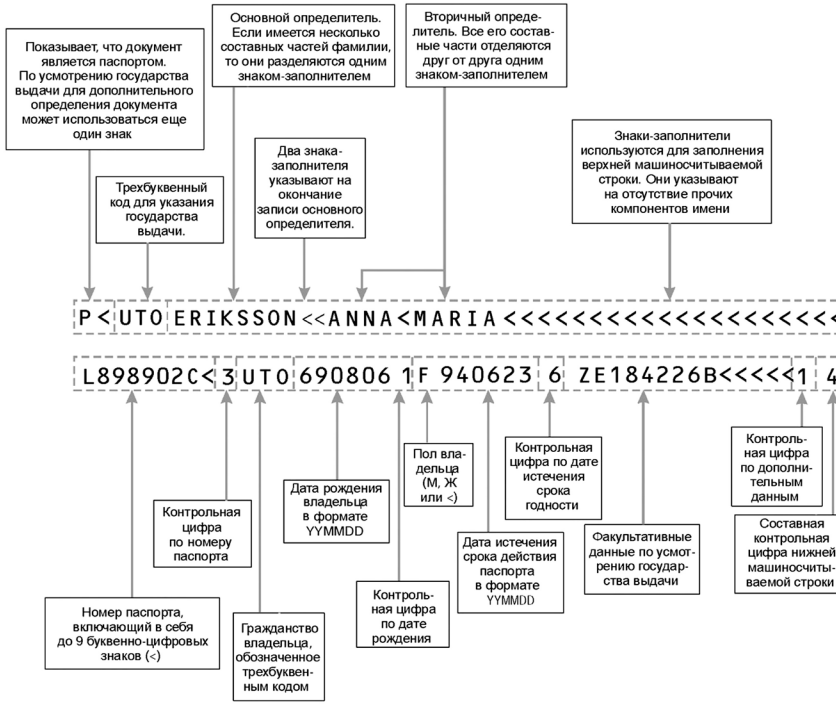

Рис. 4. Структура машиночитаемой зоны

- Дата рождения — если две цифры года рождения больше двух цифр текущего года, то это прошлый век (год вида 19XX) иначе это 20XX год.
- Дата окончания срока действия — если год вида 9X, то это прошлый век, все остальное год 20XX.

Таким образом, последовательность обработки машиночитаемой зоны выглядит следующим образом:

- проверка наличия данных в зоне;
- нахождение полей в строках согласно их позициям;
- проверка контрольных сумм, как для отдельных значений, так и для всей строки;
- приведение значений с датами к четырехзначному году.

В результате мы получим набор данных, который является корректным для документа.

Рассмотрим теперь вопрос корректировки полей зоны данных, используя значения машиночитаемой зоны.

4. Корректировка полей

Процесс корректировки полей состоит в окончательном уточнении значений и включает в себя следующие этапы:

- отделение транслитерированной части;
- коррекция по алфавиту;
- использование машиночитаемой зоны;
- обратная транслитерация.

Отделение транслитерированной части используется на полях Фамилия и Имя-Отчество. В них русскоязычное написание отделено от англоязычного прямым слешем «/», в качестве дополнительного критерия можно использовать геометрический состав поля. Если оно состоит из двух строк, то первое — всегда русскоязычное, второе — перевод. В результате из каждого поля мы получаем на выходе два — русскоязычную часть и транслитерированную.

Коррекция по алфавиту заключается в «прореживании» буквенного состава поля и его альтернатив согласно тому подмножеству символов, которое может встречаться в данном поле. Необходимость такой коррекции связана с тем, что общее поле, используемое для распознавания зоны данных, содержит в качестве алфавита все возможные символы полей, такие как цифры, русские и английские буквы. В результате поля после разделения имеют избыточное количество альтернатив букв, например русская буква «А» в качестве альтернативы содержит английскую «A», схожую по написанию. Для каждого конкретного поля можно задать его алфавит, даты содержат цифры и точку, поля с транслитерацией — только английские буквы и т. д. Пройдя по всем буквам каждого поля, оставляем в качестве альтернатив только те, которые соответствуют алфавиту.

Поля машиночитаемой зоны после распознавания и проверки контрольных сумм содержат корректный набор данных для документа. Такие данные, как фамилия и имя в транслитерации, дата рождения и дата окончания срока действия, а также номер паспорта, могут напрямую использоваться как окончательные значения полей зоны данных. Используя дату окончания срока действия загранпаспорта, можно рассчитать дату выдачи. На текущий момент паспорта выдаются сроком на 5 лет, поэтому дата выдачи — это строго дата окончания минус 5 лет. Определив пол владельца из машиночитаемой зоны, можно с большой вероятностью восстановить значение поля «Пол» в зоне данных. Если владелец женщина, то единственное значение поля это «Ж/F», если же мужчина, то возможны два варианта «М» или «М/М», точное значение можно определить исходя из размера поля, если это один символ, то «М», иначе «М/М».

Обратная транслитерация состоит в коррекции имени и фамилии на основе их значений в транслитерации машиночитаемой зоны. Для упроще-

ния и ускорения этой процедуры вначале создаются словари соответствий часто встречающихся имен и фамилий на русском языке их транслитерированным значениям. Если значение найдено в словаре, тогда нам известно точное значение на русском языке. Однако если имя или фамилия не часто встречаемы и отсутствуют в словаре, нужно установить соответствие букв и возможных альтернатив в исходном поле их значению в транслитерации. Рассмотрим такое соответствие на примере фамилии ZHURAVLEV — ЖУРАВЛЕВ (рис. 5).

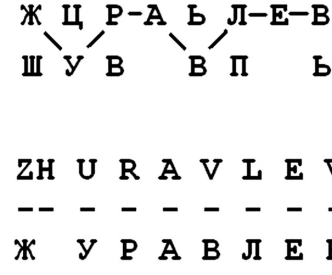

Рис. 5. Построение графа соответствий

Если рассматривать буквы как вершины, то задача сводится к построению графа, ребра которого последовательно соединяют только те из них, которые имеют правильное соответствие значениям букв в транслитерированном варианте согласно табл. 1. Нужно отметить, что помимо простого соответствия букв существуют еще правила кодирования дифтонгов — двойных гласных (подробнее можно прочитать в [2]).

В нашем примере из вариантов первой буквы только «Ж» соответствует «ZH», затем «У» соответствует «U» и т. д. В конечном итоге полученный граф дает правильное значение поля. Каждая альтернатива буквы имеет вероятностную оценку ее близости к исходному написанию. Таким образом, для каждого графа можно получить оценку его качества, например как сумму или среднее значение вероятностей входящих в него букв.

Таблица 1

Транслитерация букв

Буква	Тр.	Буква	Тр.	Буква	Тр.	Буква	Тр.
А	A	И, Й	Y	С	S	Щ	SHCH
Б	B	К	K	Т	T	Ь, Ъ	–
В	V	Л	L	У	U	Ы	Y
Г	G	М	M	Ф	F	Э	E
Д	D	Н	N	Х	RH	Ю	YU
Е, Ё	YE, E	О	O	Ц	TS	Я	YA
Ж	ZH	П	P	Ч	CH	–	–
З	Z	Р	R	Ш	SH	–	–

Если построить граф соответствия полностью не удалось, то для той буквы, которая не имеет соответствия, в качестве дополнительных альтернатив добавляем все возможные варианты сочетаний из таблицы и продолжаем граф. Поясним это на нашем примере. Допустим, первая же буква фамилии Журавлев не содержит варианта буквы «Ж», а только «Ш», которая не имеет правильного соответствия в ZHURAVLEV. Тогда, взяв из таблицы все возможные сочетания начала фамилии — это будут «Z» и «ZH», — добавляем их русские соответствия «З» и «Ж» в качестве вариантов первой буквы. В результате такого расширения можно получить несколько вариантов графов, окончательный вариант будем выбирать исходя из оценки его качества.

Когда машиночитаемая зона отсутствует или не распознана, то мы не знаем точного значения транслитерации для Имени и Фамилии. Однако это не мешает применить обратную транслитерацию и в этом случае. Основным отличием от первого варианта будет то, что графу в русском варианте будет соответствовать граф в транслитерированном, и окончательный вариант, если их несколько, будет выбираться исходя из оценки качества каждой пары.

Заключение

Отписанные методы были проверены и встроены в систему Cognitive Passport. Основным результатом является повышение качества распознавания загранпаспорта, даже в случаях отсутствия машиночитаемой зоны.

Литература

1. *Арлазаров В. Л., Постников В. В., Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Труды ИСА РАН «Управление информационными потоками». М.: URSS, 2002. С. 35–46.
2. Приложение № 7 «Правила заполнения бланков паспортов с символикой Российской Федерации» приказа МВД России от 26.05.1997 г. № 310 «Об утверждении инструкции о порядке оформления и выдачи паспортов гражданам Российской Федерации для выезда из Российской Федерации и въезда в Российскую Федерацию» (в редакции от 31.12.2003 г. № 1047). http://www.gpvu.ru/document.asp?did = 157&cid = 7&ucid = 36
3. Приложение № 11 «Правила формирования машиночитаемой зоны паспорта» приказа МВД России от 26.05.1997 г. № 310 «Об утверждении инструкции о порядке оформления и выдачи паспортов гражданам Российской Федерации для выезда из Российской Федерации и въезда в Российскую Федерацию» (в редакции от 31.12.2003 г. № 1047). http://www.gpvu.ru/document.asp?did = 159&cid = 7&ucid = 36
4. Машинносчитываемые проездные документы. Издание шестое. // Международная организация гражданской авиации. http://www.aviadocs.net/icaodocs/Docs/9303_p1_v1_cons_ru.pdf