Распознавание образов

Проблемы распознавания машиночитаемых зон с использованием малоформатных цифровых камер мобильных устройств*

К. Б. Булатов, Д. А. Ильин, Д. В. Полевой, Ю. С. Чернышова

Аннотация. Задача распознавания полученных с помощью сканирования изображений документов является классической, и в литературе описано множество методов ее решения. Распознавание машиночитаемых зон (machine-readable zone, MRZ) с камер мобильных устройств гораздо сложнее из-за слабо контролируемых условий съемки, особенностей и разнообразия камер, а также ввиду особенностей документов с машиночитаемой зоной.

В данной работе рассматривается задача распознавания MRZ-документов на фотографиях и кадрах видеопотока, полученных с камер распространенных мобильных устройств, таких как смартфоны и планшетные компьютеры. Спецификация MRZ описана в стандарте ICAO 9303 и главным образом разработана для распознавания документов с использованием специализированных сканеров и программно-аппаратных комплексов регистрации документов. В работе описаны проблемы захвата изображений MRZ-документов на естественных сценах, характерные искажения изображений, проблемы распознавания символов в условиях искажений и проблемы пост-обработки результатов распознавания с использованием языковой модели MRZ-документов.

Ключевые слова: *машиночитаемая зона, обработка изображений, оптическое распознавание символов, пост-обработка результатов распознавания, сканирование, мобильные устройства.*

Введение

Машиночитаемая зона (machine-readable zone, MRZ) является составной частью машиночитаемых документов, разрабатываемых в соответствии с международными рекомендациями, представленными в документе Doc 9303 Международной организации гражданской авиации (International Civil Aviation Organization, ICAO) [1], [2], [3]. Примером документа, изготовленного в соответствии с данными рекомендациями, является машиночитаемая зона заграничного паспорта Российской Федерации. Данный стандарт разработан с целью достичь максимального уровня автоматизации при обработке удостоверений личности при помощи программно-аппаратных инструментов с использованием методов компьютерного зрения. Распознавание MRZ-документов находит широкое применение в правительственных, финансовых, коммерческих и других областях и используется для ускорения процедур пересечения международных границ, международной унификации механизмов обработки документов, удостове-

* Исследование выполнено при частичной финансовой поддержке Российским фондом фундаментальных исследований (проекты 13–07–12172 офи_м и 15–07–06520 а).

ряющих личность, и т. п. На сегодняшний день, в условиях возрастающего интереса к более мобильным системам распознавания документов, распознавание MRZ-документов с камер мобильных устройств является интересной и актуальной задачей компьютерного зрения.

1. Распознавание отсканированных изображений машиночитаемых зон документов

Рассмотрим особенности использования сканирующего оборудования в задачах оптического распознавания документов [4]. При сканировании документ расположен в плоскости, перпендикулярной оптической оси, на фиксированном расстоянии от регистрирующей матрицы. Этим достигается гомотетичность исходного документа и его изображения, а незначительные искажения при небольших отклонениях от такого расположения легко детектируются и корректируются. В процессе сканирования документ неподвижен, поэтому исключены связанные со смещением исходного документа дефекты (размытие) изображения. Освещение в сканере формируется специальными мощными лампами подсветки, которые гарантируют стабильные характеристики освещенности и отсутствие теней.

Особым случаем сканирующего оборудования являются специализированные документные считыватели и аппаратно-программные комплексы, в которых изображение получается по принципам планшетного, планетарного или щелевого сканера. Документ в таких устройствах либо прижимается к стеклу, либо вставляется в специальную щель (рис. 1), что практически устраняет деформации сканируемой страницы документа.

Рис. 1. Примеры расположения документа при использовании специализированных считывателей, по материалам сайта www.regulaforensics.com

Такого рода считыватели позволяют получать изображения документов в различных схемах освещения (белая, инфракрасная, ультрафиолетовая, белая на просвет). При этом для оптического распознавания может использоваться схема с белым и инфракрасным освещением, которая дает высококонтрастное изображение с низким уровнем помех от фонового заполнения и элементов защиты.

Рис. 2. Пример изображений при сканировании паспорта Японии в белом и инфракрасном диапазонах, по материалам сайта www.regulaforensics.com

Известное взаимное расположение элементов освещения (лампы, светодиоды) относительно рабочей поверхности, на которой располагается документ, позволяет полностью устранить (в процессе проектирования прибора) или существенно упростить компенсацию бликов (в процессе работы).

В зависимости от модели, такого рода специализированное оборудование позволяет получать изображения в разрешении от 200 ppi и выше, при этом большинство модификаций имеет возможность получения изображений оптимального для оптического распознавания текста разрешения (300–400 ppi).

Таким образом, сканирующие устройства предоставляют изображения высокого качества с минимальными искажениями, что позволяет осуществлять оптическое распознавание текста с высокими качеством и надежностью [5], [6].

2. Распознавание полученных с помощью малоформатных цифровых камер изображений машиночитаемых зон документов

При съемке документов мобильными устройствами возникают типичные для съемки малоформатными цифровыми камерами проблемы [7]. По сравнению со сканерами, оптическая схема камеры является более сложной и сама по себе вносит больше искажений из-за аберраций, бликов и отражений внутри оптической системы. Использование фотосенсоров (матриц) и аналоговой электроники устройствами для регистрации изображений неизбежно приводит к появлению искажений изображений, которые называют цифровым шумом. Источниками цифрового шума являются сам процесс оцифровки аналогового сигнала (ошибки квантование сигнала, тепловой шум и переноса заряда на матрице) и его дальнейшее усиление. Цифровой шум заметен на изображении в виде наложенной маски из пикселей случайного цвета и яркости. Шум более заметен на однотонных участках изображения, в особенности — на темных. В отличие от сканирования, когда качественное освещение гарантировано, при съемке камерами часто возникает ситуация недостаточной освещенности, и влияние цифрового шума многократно усиливается. Еще один источник искажений — алгоритмы сжатия изображений, что особенно заметно для кадров видеопотока.

В зависимости от характеристик объектива и положения документа относительно плоскости наводки на резкость часть или все изображение документа может быть «размыто». Если из-за движения самого документа или камеры происходит смещение во время экспозиции, то появляется «смазывание», которое усиливается в условиях недостаточной освещенности.

Проективные искажения и деформации документов при съемке

В отличии от сканеров при съемке камерой сам документ расположен в произвольной плоскости относительно плоскости сфокусированного изображения. Отклонение от перпендикулярной оптической оси плоскости приводит к проективному искажению изображения документа. При незначительных углах отклонения можно распознавать машиночитаемую зону без дополнительного проективного исправления, но в общем случае необходимо оценить параметры проективного базиса и оптическое распознавание производить для проективно исправленного изображения. При этом возможны ошибки в определении параметров проективного исправления, что приводит к геометрическим искажениям изображений символов. Более того, как объект физического мира, исходный документ подвержен механическим деформациям. Например, выполненные на бумаге документы подвержены «изгибам» и «скручиванию» (чаще всего вдоль или поперек основного направления чтения), причем иногда возникают «волны», когда изгибы разнонаправленны в разных местах страницы. При съемке камерой обеспечить отсутствие деформаций такого рода сложно или просто невозможно (рис. 3).

Рис. 3. Нелинейные искажения строк машиночитаемой зоны после проективного исправления изображения документа (цитируется по [4])

Механическая деформация страницы документа комбинируется с проективным искажением изображения. Выровненные в параллельные строки на исходном документе символы на изображении даже после проективной нормализации могут не иметь базовых линии. Более того, искажению подвергаются не только сами строки, но и изображения отдельных символов. То есть даже после правильной проективной нормализации всего документа изображение символа из физически деформированной области на исходном документе будет отличаться от изображения этого же символа из недеформированной области. С примерами искаженных изображений символов после проективного исправления зоны можно ознакомиться на рис. 4.

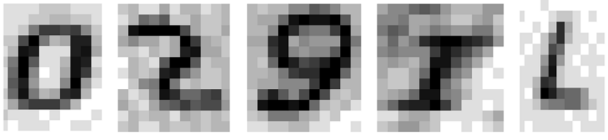

Рис. 4. Примеры искажения символов машиночитаемой зоны

Особенности

Однако, кроме общих проблем использования малоформатных камер, существует сложности, специфичные именно для документов [4]. Для машиночитаемой зоны ICAO 9303 устанавливает, что печать текста должна быть визуально разборчивой и иметь черный цвет (на длинах волн B425–B680 согласно

стандарту ИСО 1831), а так же краска должна хорошо поглощать в ближней части инфракрасного диапазона (диапазоне B900 в соответствии со стандартом ИСО 1831). Никакие защитные слои не должны отрицательно влиять на это свойство. При этом после окончательной подготовки, например, после наложения любого защитного слоя, минимальный контрастный сигнал печати (КСП/мин) после измерения в соответствии со стандартом ИСО 1831 должен соответствовать следующим параметрам: КСП/мин $\geq$ 0,6 в спектральном диапазоне B900. Таким образом, требования к контрастности налагаются только для инфракрасной области спектрального диапазона. На практике это приводит к тому, что при соблюдении стандарта некоторые страны используют для печати фонового заполнения машиночитаемой зоны краски, которые «прозрачны» в инфракрасном диапазоне, и в тоже время довольно «плотны» в оптическом (рис. 5).

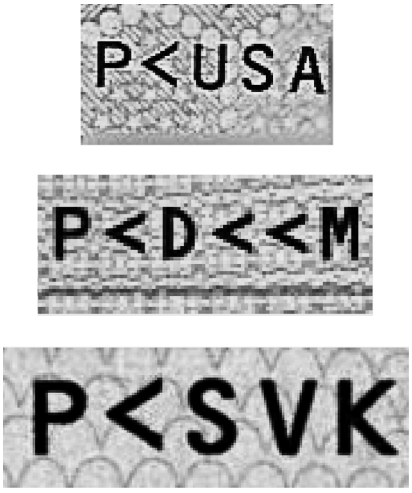

Рис. 5. Примеры фрагменты машиночитаемых зон с «плотным» в оптическом диапазоне фоновым заполнением

Для малоформатных камер мобильных устройств съемка в инфракрасном диапазоне невозможна, поэтому неоднородный фон существенно усложняет процесс оптического распознавания зоны, особенно в условия «неудачного» освещения.

Схема освещения документа в сканерах минимизирует появление теней и бликов даже для «глянцевых» страниц документа. При съемке камерой в естественных сценах на изображениях часто возникают перепады яркости (тени, отражения, рефлексы и т. д.) и цветовые искажения, которые усложняют задачи анализа изображений и распознавания, например, за счет потери существующих или появления фальшивых границ объектов. Страницы большинства документов с машиночитаемой зоной либо изготовлены из специального пластика, либо покрыты защитной пленкой и обладают хорошими отражающими свойствами. Такие физические свойства объектов съемки приводят к появлению на документе бликов (рис. 6-а). Дополнительные элементы защиты документа часто содержат области с «голографическими» элементами (рис. 6-б, 6-в), которые тоже искажают изображение.

а) б) в)

Рис. 6. Фрагмент зоны а — блик от протяженного источника света (лампа дневного света), б, в — «голографическая защита» (цитируется по [4])

3. Проблемы использования шрифта OCR-B

Рассмотрим влияние вышеперечисленных сложностей при использовании малоформатных цифровых камер на распознавание одиночных символов. Для распознавания использовались сверточные нейронные сети, обученные на проективно искаженных символах OCR-B. В качестве вектора признаков использовалось масштабированное полутоновое изображение. Для анализа использовалась эталонная база образов символов, состоящая из 300 000 изображений, полученных с помощью камер мобильных устройств. Посимвольная статистика основных типов ошибок представлена в следующей таблице.

Таблица 1

Основные классы ошибок распознавания символов шрифта OCR-B

Символ	Доля ошибок в %	Наиболее частый неверный ответ и его доля в %
«0» (цифра)	53,6 %	«O» (буква), 51,4 %
«<»	8,7 %	«2», 3,5 %
«8»	4,2 %	«B», 1,2 %
«O» (буква)	2,4 %	«0» (цифра), 2,3 %
«4»	2,3 %	«6», 0,6 %
«6»	2,2 %	«G», 0,4 %
«1»	2,0 %	«T», 0,7 %
«L»	1,7 %	«I», 1,1 %
«M»	1,7 %	«H», 0,8 %
«E»	1,7 %	«C», 0,6 %
«A»	1,7 %	«J», 0,8 %
«2»	1,5 %	«7», 1,1 %
«T»	1,4 %	«1», 0,9 %

В первом столбце располагается распознаваемый символ, во втором — доля ошибок на этом символе по отношению ко всем ошибкам в целом. В третьем столбце расположена самая частая ошибка распознавания данного символа (с ее долей).

Для печати строк текста машиночитаемой зоны ICAO 9303 устанавливает допустимое подмножество символов шрифта OCR-B, при этом некоторые символы имеют «схожие» начертания (рис. 7–11).

Наиболее трудными для различения между собой являются буква «O» и цифра «0», изображения которых отличаются только пропорциями и небольшим различием в «кривизне» (см. рис. 8).

При низком качестве съемки, особенно при появлении бликов или шумов, становятся неразличимы изображения символов тройки «M» — «N» — «H», поскольку их характерные различия ограничены небольшой центральной частью (см. рис. 9).

Небольшие искажения в виде сдвигов символов на знакоместах, размытия и линий фона могут приводить к неразличимости между собой образов символов «1» — «T» — «I» (см. рис. 10).

Стоит также отметить, что некорректное распознавание символов из пары «8» — «B» при некоторых типах искажений является обычным явлением (см. рис. 10).

Таким образом, при использовании для получения изображений документов малоформатных цифровых камер, в общем случае, невозможно гарантировать высокое качество изображения символа. Это приводит к существенно более низкому качеству и надежности результатов распознавания отдельных символов, а механизмы контекстной обработки начинают играть гораздо более важную роль (по сравнению со сканированием).

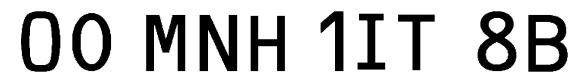

Рис. 7. Примеры «близких» в условиях искажения по графическому начертанию символов шрифта OCR-B

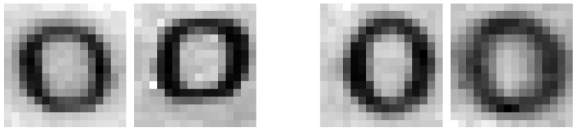

Рис. 8. Примеры сложных для распознавания символов ноль «0» (слева) и «O» (справа)

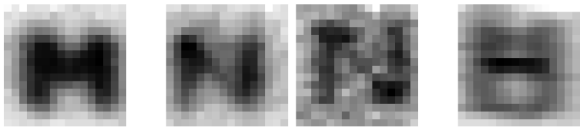

Рис. 9. Примеры сложных для распознавания символов «M» – «N» – «H»

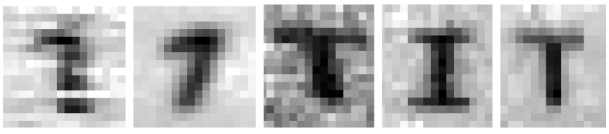

Рис. 10. Примеры сложных для распознавания символов «1» – »T» – «I»

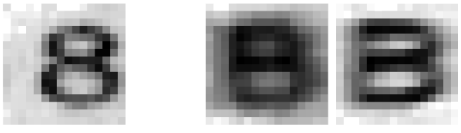

Рис. 11. Примеры сложных для распознавания символов «8» (слева) и «B» (справа)

4. Проблемы языковой модели

В современных системах распознавания и идентификации структурированных документов для улучшения точности распознавания используются механизмы статистической коррекции. Эти механизмы используют информацию о структуре документа, о «контексте» распознавания, и опираются на языковую модель распознаваемого документа (либо распознаваемого поля). [8]. Известны алгоритмы подобной статистической коррекции, или пост-обработки, опирающиеся на группу родственных методов, таких как скрытые марковские модели (Hidden Markov Models, HMM) [9], конечные автоматы, N-граммные и словарные методы, [10], а также механизмы, использующие взвешенные конечные преобразователи (Weighted Finite-State Transducers, WFST) [11].

Мощность контекста

Рассмотрим некоторое текстовое поле F. С точки зрения структуры документа поле F обладает некоторой семантической структурой. С точки зрения представления документа поле F также обладает некоторой синтаксической структурой. На основе семантики документа и синтаксической структуры представления документа для поля можно определить некоторую модель языка. К примеру, пусть F — это поле «дата рождения держателя» машиночитаемой зоны заграничного паспорта Российской Федерации. Тогда согласно семантической структуре в F содержится информация о годе, месяце и дне рождения держателя. Так как MRZ заграничного паспорта РФ выполнена в соответствии с рекомендациями ICAO 9303, в структуре данных MRZ для поля F выделена отдельная фиксированная позиция (вторая строка MRZ, символы 14–19, с контрольной суммой в 20-м символе) и для него определена синтаксическая структура: дата записывается в формате

YYMMDD, где YY — последние две десятичных цифры года, MM — десятичное представление номера месяца, DD — десятичное представление номера дня в месяце. Если какая-либо часть (или несколько частей) даты рождения неизвестна, соответствующие знакоместа отмечаются символами-заполнителями («<»). Контрольная сумма представляется в виде одной десятичной цифры и ее значение вычисляется по алгоритму, специфицированному рекомендациями ICAO 9303 [1].

На основе определенных семантической и синтаксической структур поля можно определить языковую модель, которая будет кодировать множество возможных значений поля. Представлять такую языковую модель можно несколькими способами, например при помощи БНФ-грамматики, либо в виде регулярного языка, кодируемого конечным автоматом. Один из возможных способов представления языковой модели является построение проверяющей грамматики G, состоящей из алфавита, множества всевозможных строк, составленных из символов этого алфавита, а также предиката принадлежности P. Слово S из множества всевозможных слов соответствует языковой модели G, если предикат P принимает на слове S истинное значение. Так как в документе ICAO 9303 для каждого поля предусмотрены некоторые правила, ограничивающие множество возможных значений поля (т. е. усиливающие предикат P), а также механизм контрольной суммы, представление языковой модели поля в виде проверяющей грамматики G достаточно удобно.

Задача статистической коррекции результата распознавания поля F с проверяющей грамматикой G ставится достаточно просто: во взвешенном множестве возможных альтернативных значений поля F найти значение с максимальным весом, на котором предикат P выполняется. Если количество всевозможных значений F конечно (к примеру, максимальная длина поля ограничена), можно определить «мощность контекста» как отношение мощности области ложности предиката P к мощности множества всевозможных значений F. Чем больше это отношение, тем «мощнее» контекст поля, и, соответственно, тем больше вероятность успешного исправления результата распознавания. К примеру, из всевозможных строк длины 7, состоящих из десятичных цифр, менее 0,4 % являются действительными датами (учитывая контрольную сумму), соответственно мощность контекста для этого поля превышает 99,6 %.

Поле «код MRZ-документа» машиночитаемой зоны

Поле «код документа» (document code) является двухсимвольным идентификатором типа MRZ-документа. Код документа располагается в самом начале первой строки MRZ-зоны, вне зависимости от типа MRZ-документа, и алфавит его первого символа строго зафиксирован («P» для паспортов, «V» для виз, «A», «C» или «I» для других документов, удостоверяющих личность), что позволяет построить для этого символа достаточно надежную процедуру коррекции результата распознавания. Однако второй символ кода документа оставлен на усмотрение организации, выдающей документ. Так как общая контрольная сумма на поле «код документа» не распространяется, языковую модель (помимо общего ограничения на алфавит) для второго символа типа документа в общем случае построить нельзя. Стоит также отметить, что существуют организации, осуществляющие выпуск специальных документов, которые по своей синтаксической структуре напоминают MRZ-документы, но таковыми не являются. В таких документах может встречаться первый символ кода документа, не предусмотренный стандартом ICAO 9303. Примером таких документов является MRZ-подобная зона на водительских правах Республики Молдова образца 1995–2010 годов. Структура MRZ-подобной зоны на данном типе документов совпадает со структурой документов типа TD-2, предусмотренных документом ICAO 9303 [3], за исключением поля «код документа» (рис. 12).

Рис. 12. Пример MRZ-подобного фрагмента документа, не соответствующего ICAO 9303

Поля «код выдающего органа» и «гражданство» машиночитаемой зоны

Поля «код выдающего органа» (issuing state/authority) и «гражданство» (nationality) определяют, соответственно, уникальный код организации, выдавшей документ, содержащий MRZ-зону, и гражданство держателя документа. Данные коды базируются на трехбуквенных кодах государств согласно ISO 3166-1 с некоторыми расширениями (добавлены коды, соответствующие специальным внеправительственным организациям, уполномоченным выдавать документы, удостоверяющие личность, и коды для лиц без определенного гражданства). Языковой моделью обоих полей может являться словарь — т. е. просто конечное множество всевозможных трехбуквенных кодов. Доля допустимых кодов из всевозможных трехбуквенных слов составляет ~1,4 %, следовательно мощность контекста такой языковой модели достаточно высока — ~98,6 %.

Поле «имя держателя документа» машиночитаемой зоны

Поле «имя» (name) является одним из самых сложных полей с точки зрения стандартизации, принимая во внимание разнообразие структуры имени в разных странах и в разных языках. В документе ICAO 9303 описаны некоторые требования к оформлению имени, на основе которых можно составить первичные проверочные правила: поле «имя» может состоять из одной либо двух секций, разделенных двумя символами-заполнителями («<»), каждая секция может состоять из одного или нескольких слов, разделенных одним символом-заполнителем. Каждое слово должно состоять только из букв латинского алфавита. Никаких дополнительных механизмов проверки документом ICAO 9303 не предусмотрено (общая контрольная сумма MRZ-документа на поле «имя» не распространяется). Для поля «имя» можно использовать известные методы пост-обработки подобных полей, такие как N-граммные модели и словарные методы [8, 12, 13, 14].

Поля «номер документа» и «персональный номер» машиночитаемой зоны

Поля «номер документа» (document number) и «персональный номер» (personal number / optional data) являются полями со слабо закрепленной синтаксической структурой, в связи с чем построить достаточно мощный механизм статистической коррекции затруднительно. Алфавит у этих полей не ограничен (т. е. ограничен только символами, возможными в MRZ-документе). Если для поля «номер документа» есть рекомендация, согласно которой номер не должен содержать символов-заполнителей в начале и середине номера (т. е. поле можно дополнять символами-заполнителями до нужной длины, но только в конце), то синтаксическая структура поля «персональный номер» полностью оставлена на усмотрение организации, выдающей документ. У обоих полей предусмотрена контрольная сумма, но даже с ее использованием эффективность механизма пост-обработки недостаточно высока: так как в алфавите присутствуют как буквы, так и цифры, эффективность пост-обработки падает из-за особенностей вычисления контрольной суммы по алгоритму, описанному в ICAO 9303. Мощность контекста для обоих полей можно увеличить, используя результат других полей, таких как «код выдающего органа». Некоторые организации, выдающие документы, удостоверяющий личность, определяют собственную синтаксическую структуру полей «номер документа» и «персональный номер». Соответственно, после того, как результат распознавания поля «код выдающего органа» получен и исправлен, синтаксическую структуру полей «номер документа» и/или «персональный номер» можно уточнить, если ограничения выдающей организации заранее известны.

Поля «дата рождения» и «дата окончания срока действия документа»

Синтаксическая структура полей «дата рождения» (date of birth) и «дата окончания срока действия» (date of expiry) описана выше в качестве примера к определению термина «мощность контекста». Эти поля являются самыми удачными с точки зрения языковой модели — алфавит их символов жестко зафиксирован (только цифры, за исключением отдельно рассматриваемого случая неизвестной даты) и на основе семантической структуры поля даты можно построить языковую модель с достаточно мощным контекстом. При построении алгоритма комбинированной пост-обработки нескольких полей можно также учитывать тот факт, что срок окончания действия документа не может быть раньше, чем дата рождения держателя, поэтому мощность контекста для совместного рассмотрения этих полей можно еще сильнее увеличить.

Поле «пол держателя документа»

Поле «пол держателя документа» (sex) представляет из себя один символ: «M» — мужской, «F» — женский, «<» — неопределенный/неизвестный. Ввиду ограниченности алфавита для данного поля можно построить достаточно эффективный механизм пост-обработки.

Контрольные суммы

Согласно документу ICAO 9303 контрольная сумма предусмотрена для полей «номер документа», «дата рождения», «дата окончания срока действия» и «персональный номер». Предусмотрена также так называемая «общая контрольная сумма» (composite check digit), при помощи которой происходит повторная валидация этих четырех полей, однако общая контрольная сумма предусмотрена не во всех типах документов (она отсутствует в так называемых MRV-A и MRV-B — вариантах машиночитаемых виз [2]). Контрольная сумма занимает один символ MRZ-зоны для каждого поля, и вычисляется следующим образом [1]:

1. Каждому символу поля назначается его вес. Первому символу назначается вес 7, второму — 3, третьему — 1. Четвертому — 7, пятому — 3, и т. д. циклично повторяя веса 7, 3 и 1.
2. Код каждого символа умножается на его вес. Код символа-заполнителя («<») равен нулю, код каждой десятичной цифры равен значению этой цифры, код каждой буквы латинского алфавита равен (9 + номер буквы в алфавите) (Код буквы «A» равен 10, код «B» — 11 и так далее. Код буквы «Z» равен 35).

3. Полученные произведения суммируются. Значением контрольной цифры является остаток полученной суммы по модулю 10.

Так как финальная сумма взвешенных кодов символов берется по модулю 10, возникает значительное количество коллизий. Особенные трудности вызывают коллизии на парах символов, которые трудно различимы механизмами распознавания одиночных символов в условиях распознавания с камеры мобильных устройств. Так, одинаковые коды (взятые по модулю 10) имеют символы «F» и «P», «H» и «R», «G» и «6», «S» и «8». В таких полях, как «номер документа» и «персональный номер», могут встречаться как цифры, так и буквы латинского алфавита, и основным способом валидации является именно контрольная сумма. Однако если на этапе распознавания одиночных символов один из символов из приведенных выше пар ошибочно распознался как другой член этой пары, то контрольная сумма при этом не изменится, и вероятность того, что после пост-обработки результат распознавания поля исправится, сильно понижается.

Веса, на которые умножаются коды символов проверяемого поля, также вызывают вопросы. К примеру, веса 7 и 3, применяемые к соседствующим символам, дают в сумме 10. Это значит, что стоящие рядом одинаковые символы (или разные символы, но с одинаковыми кодами по модулю 10) с весами 7 и 3 будут вместе давать нулевой вклад в контрольную сумму, вне зависимости от того, какие это символы. Это в свою очередь означает, что если на фотографии или на кадре видеопотока, на котором происходит распознавание MRZ-документа, наблюдается локальное искажение, из-за которого два соседних символа распознались с ошибкой (например, пара цифр «00» распозналась как пара букв «OO»), и эти два символа находятся в позициях поля с весами 7 и 3, то при помощи контрольной суммы их исправить не удастся. Особенно это касается полей «номер документа» и «персональный номер», так как у них самый широкий алфавит из всех полей MRZ-документа (в их записи допускаются как цифры, так и буквы).

Для повышения надежности механизма валидации чувствительных данных в документе ICAO 9303 для некоторых типов MRZ-документов предусмотрена общая контрольная сумма. Однако общая контрольная сумма распространяется не на весь MRZ документ, а только на те его поля, которые уже защищены собственной контрольной суммой.

Заключение

В данной работе проведен анализ MRZ-документов с точки зрения распознавания на мобильных устройствах. Спецификации стандарта ICAO 9303 представляются достаточно продуманными с точки зрения распознавания с использованием сканеров и специализированных считывателей документов, но по большей части ввиду особенностей этих устройств. Распознавание MRZ-документов на изображениях, полученных с камер мобильных устройств, является более трудной задачей, главным образом из-за мало контролируемых условий сьемки и из-за искажений изображений документа, которые не всегда можно исправить. Использование шрифта OCR-B создает проблемы ввиду графического сходства образов некоторых символов в условиях оптических искажений. Альтернативные семейства шрифтов с более ярко выраженными характерными особенностями отдельных символов, такие как OCR-A или SEMI M12, в этом случае подошли бы лучше.

С точки зрения пост-обработки результатов распознавания некоторые поля MRZ-документов, предусмотренные стандартом ICAO 9303, позволяют построить достаточно мощные алгоритмы контекстного исправления. Однако для таких полей, как «номер документа», «персональный номер» или «имя», определение более строгой синтаксической структуры помогло бы достичь гораздо более высокой точности и надежности распознавания.

Особенности алгоритма вычисления контрольной суммы, а так же тот факт, что общая контрольная сумма не покрывает весь MRZ-документ, усложняют процесс статистической коррекции результатов распознавания MRZ-документов, как для систем распознавания изображений, полученных с камер мобильных устройств, так и для традиционных аппаратно-программных комплексов с использованием специализированных сканеров.

Литература

1. ICAO Doc 9303 Part 1. 2006. Machine Readable Travel Documents — Machine Readable Passports. Sixth Edition. In two volumes. International Civil Aviation Organization.
2. ICAO Doc 9303 Part 2. 2005. Machine Readable Travel Documents — Machine Readable Visas. Third Edition. International Civil Aviation Organization.
3. ICAO Doc 9303 Part 3. 2008. Machine Readable Travel Documents — Machine Readable Official Travel Documents. In two volumes. International Civil Aviation Organization.
4. *Арлазаров В. В.*, *Жуковский А.*, *Кривцов В.*, *Николаев Д.*, *Полевой Д.* Анализ особенностей использования стационарных и мобильных малоразмерных цифровых видео камер для распознавания документов // Информационные технологии и вычислительные системы. 2014. № 3. С. 71–78.
5. *Llados J.*, *Lumbreras F.*, *Chapaprieta V.*, *Queralt J.* ICAR: Identity Card Automatic Reader // Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001. P. 470–474.

6. *Bessmeltsev V.*; *Bulushev E.*, *Goloshevsky N.* High-speed OCR algorithm for portable passport readers // Graphicon'11. 2011. P. 25–29.
7. *Doermann D.*, *Liang J.*, *Huiping L.* Progress in camera-based document image analysis //. Proceedings of Seventh International Conference on Document Analysis and Recognition. 2003. V. 1. P. 606–616.
8. *Шоломов Д. Л.*, *Постников В. В.*, *Марченко А. А.*, *Усков А. В.* Пост-обработка результатов OCR распознавания использующая частично определенный синтаксис // Сб. трудов ИСА РАН «Интеллектуальные информационные технологии. Концепции и инструментарий». М.: КомКнига/URSS, 2005. Т. 16. С. 146–165.
9. *Bouchaffra D.*, *Govindaraju V.*, *Srihari S. N.* Postprocessing of Recognized Strings Using Nonstationary Markovian Models // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997. V. 21. № 10. P. 990–999.
10. *Kukich K.* Techniques for Automatically Correcting Words in Text. ACM computing survey, Computational Linguistic 1992. V. 24, N 4, P. 377–439.
11. *Llobet R.*, *Cerdan-Navarro J.-R.*, *Perez-Cortes J.*, *Arlandis J.* «OCR Post-processing Using Weighted Finite-State Transducers». Pattern Recognition (ICPR), 2010. P. 2021–2024.
12. *Шоломов Д. Л.* Коррекция распознанного текста с использованием методов классификации. // Сб. трудов ИСА РАН. 2007. Т. 17, С. 352–366.
13. *Кляцкин В. М.*, *Котович Н. В.*, *Славин О. А.* Определение расстояния между словами в алгоритмах словарной корректировки результатов распознавания. // Технология програм. и хран. данных. В сб.: Труды ИСА РАН / под ред. В. Л. Арлазарова, Н. Е. Емельянова. М.: Ленанд/URSS, 2009. С. 260–266.
14. *Славин О. А.*, *Янишевский И. М.*, *Богданова Е. Б.* Алгоритм подтверждения результатов распознавания с помощью словаря // Инф. технол. и выч. сис. 2011. № 3.

Булатов Константин Булатович. М. н. с. ИСА РАН, аспирант НИТУ МИСиС. Окончил НИТУ МИСиС в 2013 г. Количество печатных работ: 3. Область научных интересов: распознавание образов, машинное обучение, информационные системы, обработка изображений. E-mail: hpbuko@gmail.com

Ильин Дмитрий Алексеевич. Аспирант ИСА РАН. Окончил МФТИ в 2014 г. Количество печатных работ: 2. Область научных интересов: компьютерное зрение и машинное обучение. E-mail: dmitry.ilin@phystech.edu

Полевой Дмитрий Валерьевич. С. н. с. ИСА РАН, к. т. н. Окончил МФТИ в 2004 г. Количество печатных работ: 11. Область научных интересов: методы искусственного интеллекта, оптическое распознавание, системы обработки документов. E-mail: dvpsun@gmail.com

Чернышова Юлия Сергеевна. Студентка НИТУ МИСиС. Область научных интересов: компьютерное зрение, системы распознавания документов. E-mail: chernyshovayulia07@gmail.com