Методы пост-обработки результатов распознавания машиночитаемой зоны документов*

О.О. Петрова[I,III], К.Б. Булатов[I,II,III]

[I] Московский физико-технический институт (государственный университет), г. Москва, Россия
[II] Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия
[III] ООО «Смарт Энджинс Сервис», г. Москва, Россия

Аннотация. В работе рассматривается задача коррекции (пост-обработки) результатов распознавания машиночитаемой зоны документов (MRZ). Проводится обзор существующих методов исправления ошибок распознавания и предлагается алгоритм их применения для пост-обработки MRZ. Приведены результаты экспериментальной проверки предложенных методов.
Ключевые слова: *распознавание документов, MRZ, коррекция результатов распознавания.*
DOI: 10.14357/20790279180505

Введение

Большая часть современных документов, удостоверяющих личность, таких как паспорта и идентификационные карточки, имеют машиночитаемую зону (Machine Readable Zone, MRZ), информация, содержащаяся в которой частично дублирует информацию из зоны визуальной проверки. На рис. 1 приведен пример документа, содержащего машиночитаемую зону.

Рис. 1. Пример документа с MRZ

Изначально документы с MRZ были введены в обращение с целью упрощения и ускорения проверки пассажиров сотрудниками служб паспортного контроля. Для распознавания данных использовались специальные сканеры, однако на сегодняшний день сфера применения автоматизированного ввода данных расширилась и возникла потребность в распознавании документов без использования специального стационарного оборудования, что наложило новые требования на качество распознавания, способность исправления ошибок и быстродействие систем оптического распознавания (OCR-систем).

Процесс распознавания данных машиночитаемой зоны, как правило, предполагает несколько этапов [1, 2], показанных на рис. 2:

Рис.2. Этапы распознавания MRZ

* Работа выполнена при частичной финансовой поддержке РФФИ (проекты №№ 17-29-03236, 17-29-03370).

В данной работе рассматривается этап пост-обработки результата распознавания MRZ.

1. Методы коррекции результата распознавания

Задача распознавания символов может рассматриваться как задача классификации. Результатом классификации символа является вектор альтернатив:

$$C(x) = \overline{a}\left(a_1, a_2, \ldots, a_K\right), \quad a_m = \left(i_m, g_m\right),$$

$$i_m \in \left\{1, 2, \ldots, K\right\}, \quad g_m \in \left[0, 1\right],$$

где $x \in X$ — образ символа, $\overline{a}$ — вектор альтернатив, i_m — индекс класса, g_m — оценка принадлежности символа x к данному классу.

Результатом распознавания поля является последовательность распознанных символов, составляющих это поле. Методы коррекции строятся на поиске варианта, удовлетворяющего известным правилам для данного поля, с наибольшей агрегированной оценкой.

Очевидно, что результат распознавания поля зависит не только непосредственно от результатов классификации каждого отдельного символа, но и от результата предыдущих этапов обработки документа, таких как сегментация текстовой строки. Однако алгоритмы пост-обработки, использующие информацию о различных альтернативных вариантах сегментации, требуют больших вычислительных затрат [3]. Поскольку, согласно стандарту ICAO Doc 9303, машиночитаемая зона печатается моноширинным шрифтом OCR-B [4], то для решения задачи коррекции результата распознавания MRZ можно не учитывать различные варианты сегментации.

Проверка корректности и исправление предварительного результата распознавания в OCR-системах может производиться основываясь на синтаксических и грамматических правилах, поиске ближайшего варианта в словаре (например, по метрике Левенштейна) и т.п. Также используются различные алгоритмы, основанные на применении статистических моделей языка (к которым можно отнести скрытые марковские модели [5] и N-граммные словари [6]). Требования к быстродействию системы и к режиму использования аппаратных ресурсов приводят к появлению сложно структурированных комбинированных методов узкой направленности, применяемых в специфических системах, таких как представленный в работе [7] метод, использующий внешние вызовы алгоритма поиска системы Google с функцией подсказки правописания.

В силу большого разнообразия документов, имеющих машиночитаемую зону, интерес вызывают методы, не привязанные к конкретному типу документа или полю, а адаптируемые в различных случаях при помощи параметризации. Одним из таких методов является описанный в [8] алгоритм, основанный на проверяющих грамматиках. Его суть заключается в эффективном переборе результатов распознавания поля до тех пор, пока предикат допустимости с точки зрения языковой модели не примет истинное значение. Поскольку одним из ключевых требований к OCR-системам является быстродействие, то для сокращения возможных вариантов значения поля, чей допустимый алфавит меньше, чем общий алфавит распознавания, можно перед применением проверяющих грамматик провести фильтрацию альтернатив для каждого символа, оставив только альтернативы, соответствующие допустимому алфавиту распознавания, тем самым исключив из перебора заведомо неверные варианты.

За счет того, что в документе ICAO Doc 9303 (или иных нормативных документах) зафиксированы синтаксические и семантические правила заполнения, для этих полей можно повысить качество распознавания путем применения алгоритмов коррекции, основанных на проверке соответствия результата распознавания шаблону.

Распознавание MRZ на мобильных устройствах в условиях аппаратных ограничений приводит к необходимости создания эффективных методов коррекции, отвечающих требованиям быстродействия, автономности и надежности распознавания.

2. Коррекция полей MRZ

Алфавит распознавания машиночитаемых зон содержит 37 символов: 10 цифр, 26 букв латинского алфавита и символ-заполнитель ‘<’.

В стандарте ICAO Doc 9303 [4] описано пять типов машиночитаемых документов, определяемых количеством строк MRZ, их длиной и разбиением на отдельные поля:

1. MRVA (Machine Readable Visa type A) – машиночитаемые визы типа A. MRZ содержит две строки по 44 символа в каждой;
2. MRVB (Machine Readable Visa type B) – машиночитаемые визы типа B. MRZ содержит две строки по 36 символов в каждой;
3. TD1 (Size 1 Machine Readable Official Travel Documents) – машиночитаемые идентификаци-

онные карточки первого типа. MRZ содержит три строки, каждая длиной 30 символов;
4. TD2 (Size 2 Machine Readable Official Travel Documents) – машиночитаемые идентификационные карточки второго типа. MRZ содержит две строки, каждая длиной 36 символов;
5. TD3 (Size 3 Machine Readable Travel Documents) или MRP (Machine Readable Passports) – машиночитаемые паспорта. MRZ содержит две строки по 44 символа в каждой.

Помимо типов, описанных в ICAO Doc 9303, существуют машиночитаемые документы, не подчиняющиеся этому стандарту.

В связи с тем, что правила для MRZ заданы в стандартах в терминах полей (то есть заданы правила заполнения по каждому полю отдельно, а не всей зоны), то для коррекции результата распознавания MRZ имеет смысл предварительное разбиение распознанных строк MRZ на отдельные поля. Для этого прежде всего требуется определить тип распознаваемого документа. При этом можно опираться на соответствие результата распознавания шаблону типа: на основе допустимых алфавитов распознавания для каждого символа каждого поля можно составить шаблон, учитывающий также количество и длину строк. На рис. 3 приведены

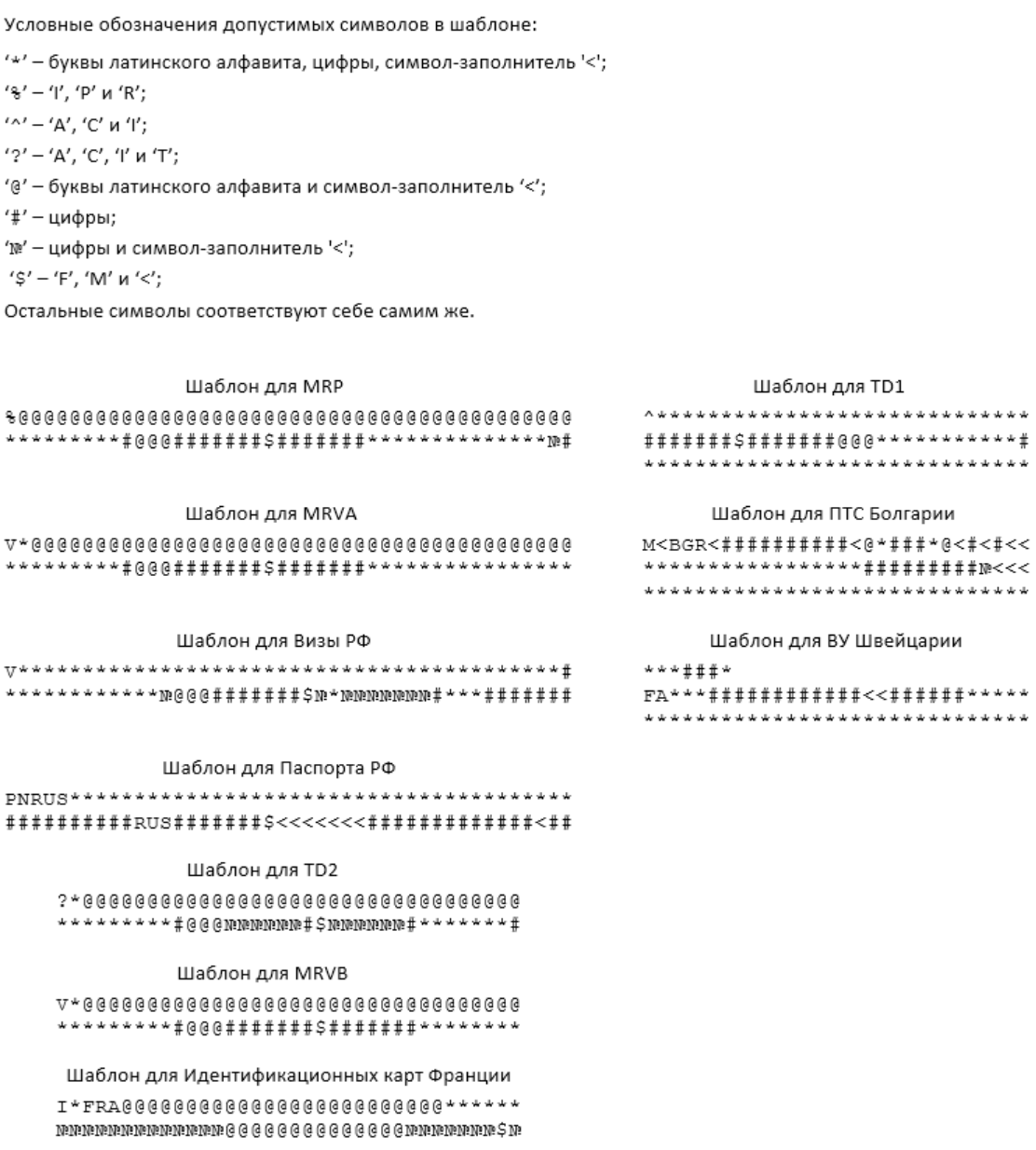

Рис. 3. Шаблоны для разных типов документов

шаблоны для пяти типов документов, соответствующих стандарту ICAO Doc 9303, и пяти нестандартных типов документов.

В общем виде процесс пост-обработки результата распознавания MRZ представлен в виде блок-схемы на рис. 4.

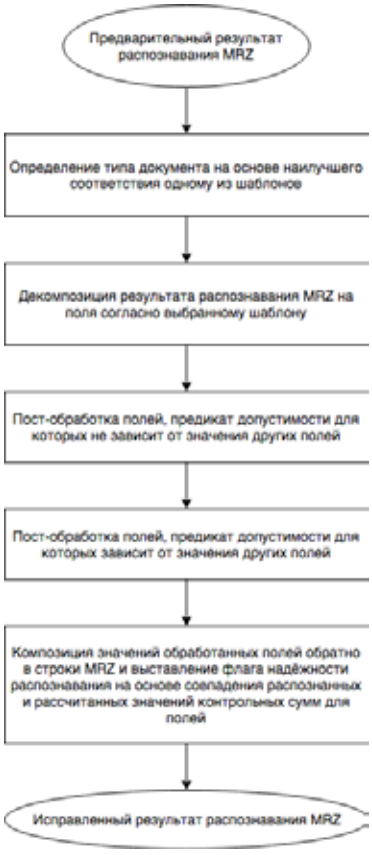

Рис. 4. Процесс пост-обработки результата распознавания MRZ

Из-за того, что предикат допустимости для некоторых полей зависит от значения других полей, при коррекции важен порядок пост-обработки полей. К примеру, предикат допустимости для полей ‘код документа’ и ‘номер документа’ может зависеть от локальных стандартов выпускающих организаций. Поэтому коррекция этих полей должна проводится после того, как возможные ошибки распознавания поля ‘код страны-эмитента’ уже исправлены.

На рис. 5 детально показаны этапы пост-обработки полей для документов типа MRP. Для других типов процесс аналогичен.

Пунктирными линиями обозначены зависимости предикатов допустимости одних полей от значений других полей.

Все допустимые значения полей ‘код страны-эмитента’ и ‘гражданство’ зафиксированы в документе ICAO Doc 9303. Поэтому корректность результата распознавания можно проверять по вхождению в словарь. Для документов, в которых заранее известна страна-эмитент и поэтому словарь содержит единственное значение, вместо использования проверяющих грамматик можно явно выписать исправленный результат распознавания. Аналогично предикат допустимости для поля ‘код документа’ может быть построен как вхождение в словарь.

Поле ‘имя держателя документа’, согласно стандарту ICAO Doc 9303, может содержать только буквы латинского алфавита и символ ‘<’, правила расстановки которого описаны в документе [4]. Поэтому предикат допустимости для этого поля учитывает только соответствие результата распознавания допустимому алфавиту и корректность расставления знака-заполнителя ‘<’. В реальных системах распознавания, когда требования к быстродействию не позволяют универсальным методам пост-обработки производить перебор больше некоторого заданного числа вариантов, грубые ошибки распознавания (к примеру, неверное распознавание символов-заполнителей, дополняющих поле до нужной длины) имеет смысл исправлять до применения универсальных методов пост-обработки.

После приведения предварительного результата распознавания поля ‘имя держателя документа’ к соответствию задаваемым стандартами правилам, возможные оставшиеся ошибки распознавания в этом поле могут быть исправлены при помощи статистических методов, например, N-грамм [6].

Возможные значения поля ‘пол держателя документа’ в MRZ, согласно стандарту ICAO Doc 9303, ограничиваются всего тремя значениями – ‘M’, ‘F’ и ‘<’. Поэтому результатом коррекции результата распознавания данного поля будет значение с максимальной оценкой, являющееся допустимым.

Мощным инструментом, позволяющим использовать проверяющие грамматики для коррекции результата распознавания MRZ, являются предусмотренные для некоторых полей контрольные суммы. Алгоритм их вычисления описан в документе [4].

Для полей ‘дата рождения держателя документа’ и ‘дата окончания срока действия документа’ предикат допустимости, помимо контрольной суммы, учитывает семантическую корректность даты (значение месяца от 1 до 12, допустимое количество дней в месяце).

Поле ‘номер документа’ также верифицируется контрольной суммой. Кроме того, для некоторых документов существуют стандарты, пропи-

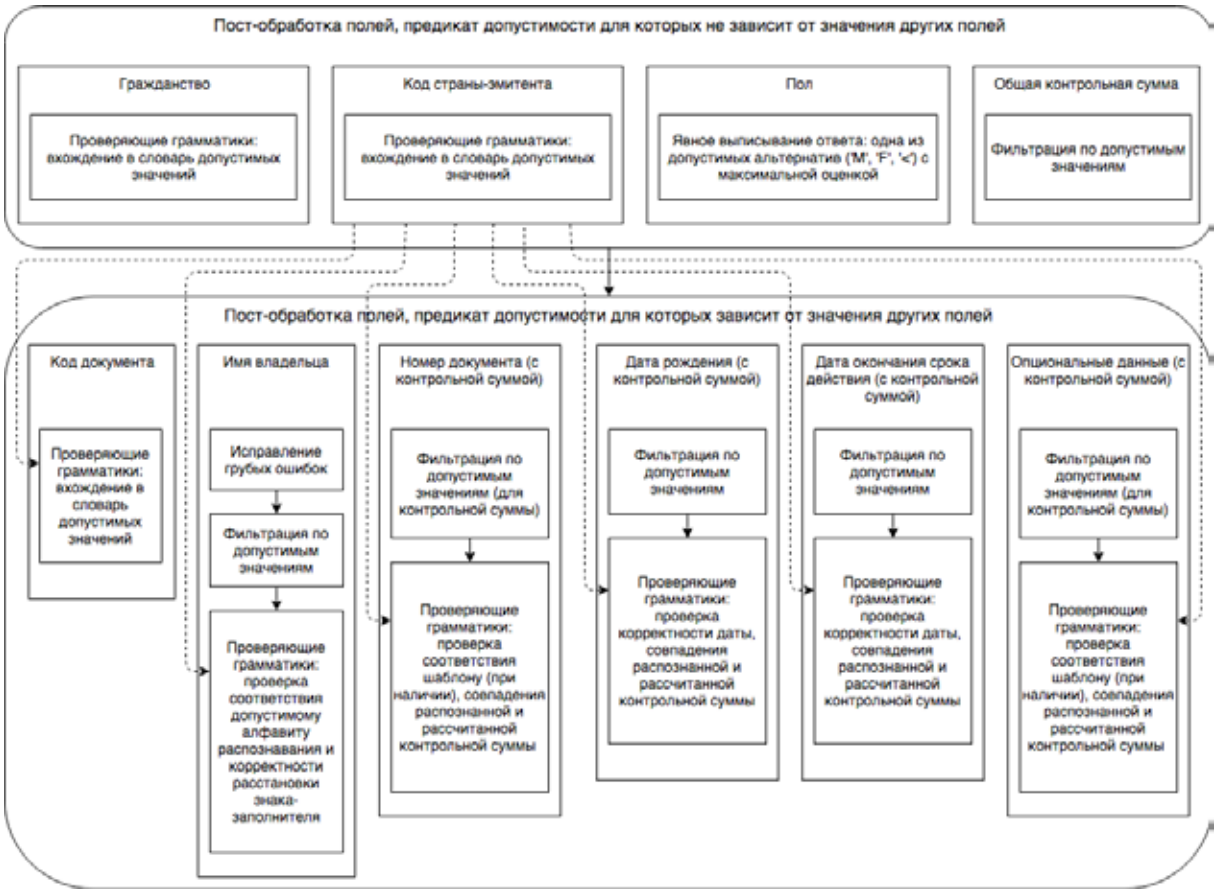

Рис. 5. Пост-обработка полей MRP

санные в документах выпускающих организаций, позволяющие проверять соответствие номера документа некоторому шаблону, зависящему от страны-эмитента. Аналогично поля ‘опциональные данные’, ‘номер приглашения’ (для российских виз) исправляются, основываясь на совпадении рассчитанной и распознанной контрольной суммы (где она для корректируемого поля предусмотрена) и на соответствии шаблонам, зависящих от страны-эмитента.

При корректировке полей, для которых предусмотрена контрольная сумма, встает вопрос, считать ли контрольную сумму частью поля, то есть должно ли ее значение исправляться наравне со значением символов поля, или контрольная сумма должна использоваться исключительно для верификации результата распознавания. Поскольку поля контрольных сумм не являются значащими, то есть в большинстве OCR-систем результат распознавания контрольных сумм является лишь вспомогательной информацией для верификации данных других полей, то лучше строить общий предикат допустимости для поля и контрольной суммы, его верифицирующей. Однако в случае отличия значения контрольной суммы от ее предварительного результата распознавания, предлагается его не исправлять, а поставить флаг недостоверности распознавания поля (для визуальной проверки результата оператором). Стоит отметить, что до коррекции поля, верифицируемого контрольной суммой, стоит заранее отсечь варианты заведомо неверного распознавания контрольной суммы путем фильтрации недопустимых значений (контрольная сумма не может быть буквой). При таком способе коррекции поля с одной стороны используется свойство контрольной суммы как мощного инструмента коррекции результата распознавания, с другой – не теряется свойство быть средством верификации.

Для некоторых документов предусмотрена общая контрольная сумма, рассчитываемая по значениям нескольких полей. Поскольку в случае несовпадения распознанной и рассчитанной общей контрольной суммы локализация и исправление ошибки распознавания требует больших вычислительных затрат, то в этом случае корректировка не производится, а только выставляется флаг недостоверности распознавания всей MRZ.

Табл. 1

Результаты работы алгоритмов пост-обработки для некоторых полей MRZ.

Название поля	Количество ошибок без использования методов коррекции		Количество ошибок с использованием методов коррекции		Количество ошибок, привнесенных методами коррекции	
	Полей	Символов	Полей	Символов	Полей	Символов
MRZ целиком	695 (16,8526%)	1541 (0,4453%)	170 (4,1222%)	395 (0,1141%)	0	9 (0,0026%)
Имя владельца	283 (6,8623%)	488 (0,3218%)	125 (3,0310%)	250 (0,1649%)	0	0
Номер документа	150 (3,6372%)	202 (0,5413%)	18 (0,4365%)	39 (0,1045%)	1 (0,0242%)	4(0,0107%)
Дата рождения	144 (3,4918%)	193 (0,7800%)	7 (0,1697%)	20(0,0808)	0	1 (0,0040)
Код документа	63 (1,5276%)	67 (0,8191%)	10 (0,2425%)	11 (0,1345%)	0	0
Пол	53 (1,2852%)		1 (0,0242%)		0	

3. Общие проблемы пост-обработки машиночитаемой зоны

Для некоторых типов документов построенные на основе правил заполнения полей шаблоны очень похожи друг на друга. Например, для MRP и MRVA и для TD2 и MRVB они отличаются только в нескольких символах (см. рис. 3), что при наличии определенных ошибок может привести к неправильному определению типа документа и, как следствие, к некорректной пост-обработке.

Нарушения стандартов, которые можно разделить на систематические и единичные, также вызывает трудности при применении методов пост-обработки. Систематические нарушения стандарта (некоторые из которых приведены в работе [9]) предполагают, что система распознавания документов будет корректно их обрабатывать, однако учет подобных особенностей усложняет построение предиката допустимости. Несистематические нарушения (к которым относятся использование не предусмотренных стандартом символов, неправильно рассчитанные контрольные суммы, некорректная расстановка символа-заполнителя ‘<’ и другие) не могут быть предусмотрены при построении методов коррекции. Из-за этого методы пост-обработки могут привнести ошибку даже в изначально верно распознанный документ, пытаясь привести его в соответствие со стандартом.

Проверка контрольных сумм является хорошим способом детектировать и исправлять ошибки, однако способ их расчета приводит к коллизиям, описанным подробно в работе [10].

4. Экспериментальные данные

Работа описанных методов пост-обработки проверялась на наборе 4124, правильно разбитых на отдельные символы и не нарушающих требования стандартов в изображениях MRZ шести типов: MRP, MRVB, Паспорт РФ, Виза РФ, TD1 и TD2, полученных при помощи камеры мобильных устройств и подверженных различным искажениям [10, 11]. Распознавания изображений отдельных символов производилось при помощи сверточных искусственных нейронных сетей [12]. В табл. 1 отражены полученные результаты по пяти полям, присутствующим в документах всех типов и для всей MRZ целиком. Приведено количество ошибок с использованием и без использования алгоритмов пост-обработки.

Из полученных результатов следует, что использование предложенного способа корректировки результатов распознавания полей MRZ помогает существенно сократить количество ошибок при этом практически не привнеся новых. Большинство примеров, на которых алгоритм корректировки ухудшает изначально верный результат распознавания, делится на две группы:

1. Неверное исправление происходит в поле, для которого предусмотрена контрольная сумма. Из-за того, что алгоритм расчета контрольных сумм допускает наличие нескольких вариантов

последовательностей символов, имеющих одинаковую контрольную сумму, в случае плохого качества оцифровки распознаваемого документа при коррекции может быть изменен символ, изначально распознанный верно.

2. Ошибка корректировки произошла в случае неверного выбора типа документа и, как следствие, применения неверных шаблонов при исправлении. Например, один из документов типа MRVB был ошибочно отнесен к типу TD2, из-за чего один символ поля ‘номер документа’ был заменен на неверный.

Заключение

Была рассмотрена задача пост-обработки результатов машиночитаемой зоны документов. На основе проведенного обзора существующих методов коррекции результатов распознавания текстовых полей был предложен алгоритм исправления ошибок распознавания полей MRZ согласно синтаксическим и семантическим правилам, зафиксированным в стандартах для этих полей. Были описаны проблемы, возникающие при применении предложенного алгоритма пост-обработки. Для оценки эффективности работы описанного алгоритма коррекции использовался датасет, включающий изображения документов шести различных типов, подверженных различным искажениям. Проведенные эксперименты показали снижение количества ошибок.

В дальнейшем планируется усиление методов пост-обработки путем добавления алгоритмов коррекции, основанных на применении статистических моделей языка для полей, значения которых могут быть описаны подобными моделями, например, для поля ‘имя держателя документа’.

Литература

1. *Bessmeltsev V., Bulushev E., Goloshevsky N.* High-speed OCR algorithm for portable passport readers// Graphicon’11.2011. P. 25–29.
2. *Слугин Д.Г.* Особенности контекстного распознавания российского заграничного паспорта в системе Cognitive Passport // Труды ИСА РАН. Т. 45. 2009. С.174-182.
3. *Шоломов Д.Л., Постников В.В., Марченко А.А., Усков А.В.* Пост-обработка результатов OCR-распознавания, использующая частично-определенный синтаксис. // Труды ИСА РАН. Т. 16. 2005. С. 146-163.
4. *ICAO Doc 9303 Part 3*: Specifications Common to all MRTDs. 2015. Machine Readable Travel Documents – International Civil Aviation Organization.
5. *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.
6. *Kukich K.* Techniques for Automatically Correcting Words in Text // ACM computing survey, Computational Linguistics. 1992. V. 24. № 4, P. 377–439.
7. *Youssef Bassil, Mohammad Alwani*. OCR post-processing error correction algorithm using Google’s online spelling suggestion. // Journal of Emerging Trends in Computing and Information Sciences, 3(1):90–99.
8. *Булатов К.Б., Николаев Д.П., Постников В.В.* Универсальный алгоритм пост-обработки результатов распознавания на основе проверяющих грамматик // Труды ИСА РАН. Т. 65. Выпуск 4. 2015. С. 68-73.
9. *Mercer J.* Errors in travel documents. // Keesing Journal of Documents& Identity, issue 34, Feb 2011. URL http://keesingjournalofdocuments.com/content/Cases_analysed/KJDI_2011_34_Mercer.pdf (дата обращения 22.06.2018).
10. *Булатов К.Б., Ильин Д.А., Полевой Д.В., Чернышова Ю.С.* Проблемы распознавания машиночитаемых зон с использованием малоформатных цифровых камер мобильных устройств // Труды ИСА РАН. Т. 65. Выпуск 3. 2015. С. 85-93.
11. *Арлазаров В.В., Жуковский А.Е., Кривцов В.Е., Николаев Д.П., Полевой Д.В.* Анализ особенностей использования стационарных и мобильных малоразмерных цифровых видео камер для распознавания документов // Труды ИСА РАН. Т. 64. Выпуск 3. 2014. С. 71-81.
12. *Dong Xiao Ni.* Application of Neural Networks to Character Recognition // Proceedings of Students/ Faculty Research Day, CSIS, Pace University, May 4th, 2007.

Петрова Ольга Олеговна. Московский физико-технический институт (государственный университет), г. Москва, Россия. Студентка 2-го курса магистратуры. Область научных интересов: информационные технологии, системы распознавания. E-mail: opetrova@smartengines.biz

Булатов Константин Булатович. Институт системного анализа Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, Россия. Программист I-й категории. Количество печатных работ: 14. Область научных интересов: машинное обучение, компьютерное зрение, системы распознавания, информационные технологии. E-mail: hpbuko@gmail.com

Methods of machine-readable zone recognition results post-processing

O.O. Petrova[I,III], K.B. Bulatov[I,II,III]

[I] Moscow Institute of Physics and Technology (State University), Moscow, Russia

[II] Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

[III] LLC "Smart Engines Service", Moscow, Russia

Abstract. In this paper a task of correction (post-processing) of machine-readable zone recognition results is discussed. A survey is presented for existing approaches of recognition error correction methods and an algorithm is proposed for applying these methods for machine-readable zone post-processing. Experimental results are shown for the described methods.

Keywords: *documents recognition, MRZ, recognition results correction.*
DOI: 10.14357/20790279180505

References

1. *Bessmeltsev V., Bulushev E., Goloshevsky N.* High-speed OCR algorithm for portable passport readers// Graphicon'11. 2011. P. 25–29.
2. *Slugin D.G.* Properties of context-based recognition of Russian international passport in Cognitive Passport. // Proc. ISA RAS, V. 45, 2009. P.174-182.
3. *Sholomov D.L., Postnikov V.V., Marchenko A.A., Uskov A.V.* Post-processing of OCR Results Using Automatically Constructed Partially Defined Syntax. // Proceedings of the Institute for System Analysis RAS, Vol. 16. pp. 146-163, 2005.
4. *ICAO Doc 9303 Part 3*: Specifications Common to all MRTDs. 2015. Machine Readable Travel Documents – International Civil Aviation Organization.
5. *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.
6. *Kukich K.* Techniques for Automatically Correcting Words in Text // ACM computing survey, Computational Linguistics. 1992. V. 24. № 4, P. 377–439.
7. *Youssef Bassil, Mohammad Alwani*. OCR post-processing error correction algorithm using Google's online spelling suggestion. // Journal of Emerging Trends in Computing and Information Sciences, 3(1): 90–99.
8. *Bulatov K.B., Nikolaev D.P., Postnikov V.V.* General-purpose algorithm for text field OCR result post-processing based on validation grammars // Proceedings of the Institute for System Analysis RAS, Vol. 65(4). pp. 68-73, 2015.
9. *Mercer J.* Errors in travel documents. // Keesing Journal of Documents& Identity, issue 34, Feb 2011. Available at: http://keesingjournalofdocuments.com/content/Cases_analysed/KJDI_2011_34_Mercer.pdf (accessed June 22, 2018).
10. *Bulatov K.B., Polevoy D.V., Ilin D.A., Chernyshova Y.S.* Problems of machine-readable zone recognition captured with digital mobile cameras//Proceedings of the Institute for System Analysis RAS, Vol. 65(3). pp. 85-93, 2015.
11. *Arlasarov V.V., Zhukovsky A.E., Krivtsov V.E., Nikolaev D.P., Polevoy D.V.* Analysis of features of the use of fixed and mobile small-sized digital video camera for OCR // Proceedings of the Institute for System Analysis RAS, Vol. 64(3). pp. 71-81, 2014.
12. *Dong Xiao Ni.* Application of Neural Networks to Character Recognition // Proceedings of Students/ Faculty Research Day, CSIS, Pace University, May 4th, 2007.

O.O. Petrova. Moscow Institute of Physics and Technology (State University), Moscow, Russia Moscow. 2-nd year Master's program student. Scientific interests: information technologies, recognition systems.
E-mail: opetrova@smartengines.biz

K.B. Bulatov. Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia. I-st category programmer. Number of publications: 14. Scientific interests: machine learning, computer vision, recognition systems, information technologies.
E-mail: hpbuko@gmail.com