Применение дескрипторов объектов для привязки структурных элементов зашумленных образов деловых документов*

О. А. Славин [I,II]

[I] Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", г. Москва, Россия
[II] ООО «Смарт Энджинс Сервис», г. Москва, Россия

Аннотация. Рассматривается задача извлечения из распознанного образа документа элементов заполнения (полей) с помощью дескрипторов – описаний одного или нескольких структурных элементов. Структурными элементами могут быть слова статического текста и линии разграфки, используемые для оформления дизайна документа. Рассматриваются деловые документы с упрощенной структурой и ограниченным словарем. Рассматриваются гибкие деловые документы, допускающие существенные модификации дизайна страницы. Дескрипторы создаются с учетом значительного числа возможных ошибок распознавания страниц документов. Описываются комбинированные дескрипторы, состоящие из нескольких термов и отрезков линий. Приводится алгоритм привязки, базирующийся на дескрипторах. Экспериментально показывается, что извлечение комбинированных дескрипторов улучшает точность распознавания полей документа при распознавании на 17%, а точность извлечения информации из образа документа – на 16%. В качестве OCR в эксперименте использовалась система SDK Smart Document Engine.

Ключевые слова: зашумленный образ, распознавание документа, текстовая особая точка, дескриптор.

DOI 10.14357/20718632220402

Введение

Задача распознавания изображений документов является актуальной из-за того, что количество напечатанных на бумаге документов растет с каждым годом. Например, объемы потоков входящих и исходящих документов в крупных организациях могут достигать $O(10^6)$ страниц в день. Количество документов в архиве крупного банка может достигать $O(10^{10})$ листов. Парадоксально, что вместе с развитием электронного документооборота растет число бумажных документов и печатных копий электронных документов. Известно, что объем бумажного документооборота растет в РФ более чем на 10% в год [1]. Это в основном объясняется несовершенством систем электронного делопроизводства и документооборота. Однако рост бумажного документооборота требует не только хранения оригиналов в архиве, но и хранение оцифрованных страниц документов в электронном архиве. Ввод оцифрованных документов в информационные системы сопровождается заполнением карточек документов, содержащих

* Работа выполнена при частичной финансовой поддержке РФФИ (проект 20-07-00934 “Development, properties study and justification of anytime algorithms for computed tomography”).

атрибуты для последующего поиска. Эффективным способом извлечения атрибутов является распознавание образов страниц [2].

В работе рассматриваются деловые документы. Этот вид документов характеризуется относительно простой структурой и ограниченным словарем статических текстов. Идеально оцифрованные деловые документы могут быть распознаны с малым числом ошибок, что позволяет оптимизировать процесс создания карточек документов. Проблемой является невозможность во многих случаях организовать не только идеальное качество оцифровки, но даже контроль процесса оцифровки. Это приводит к появлению большого числа ошибок распознавания, откуда следует необходимость применения ручного труда для проверки результатов распознавания. Другой проблемой является необходимость ввода документов, структурные элементы которых не имеют фиксированного расположения на бумаге. Такие документы называются гибкими. Они создаются на основании одного шаблона или прототипа, но в местах создания документов этот шаблон подвергается модификации. В работе рассматриваются распознанные зашумленные изображения документов, в которых зашумлению подвергаются как образы символов, так и графических объектов. Эта тематика в настоящее время является актуальной [3, 4].

Документ определяется как совокупность структурных элементов, которые являются либо полями, либо статическими элементами. Статическими элементами, прежде всего, являются слова статического текста, линии таблиц различной сложности, линии подчеркивания, заготовки для пометок (чек-боксов). К особенностям деловых документов относится малый объем словаря допустимых ключевых слов, а также малое число возможных совокупностей графических элементов. Поля определяются как объект, который ограничен несколькими статическими элементами. Поле может содержать различную информацию:

- печатный или рукописный текст;
- рукописную пометку, например, подпись;
- пометку;
- графический элемент, например, печать;
- штрих-код.

Рассмотрим схему распознавания образа страницы документа, состоящую из следующих этапов:

- обработка и нормализация страницы;
- распознавание слов, извлечение графических примитивов и других объектов;
- ректификация объектов и поиск локальных особенностей;
- поиск границ полей с помощью границ локальных особенностей;
- извлечение содержимого полей в найденных границах с помощью атрибутов полей.

В данной работе будет рассмотрен метод, который позволяет определять границы полей с помощью комбинированных дескрипторов, использующих описания как текстовых, так и графических элементов оформления бланка документа. Исходными данными метода будут распознанные деловые документы, а результатом – границы полей для последующего уточненного распознавания и анализа.

Постановка задачи распознавания состоит в следующем. На основании распознавания текстовых объектов и найденных графических элементов извлечь информацию из областей, соответствующих границам полей. В работе рассматривается двухпроходный способ распознавания документа с использованием описания полей. На первом проходе после распознавания текста и извлечения графических примитивов прогнозируются границы полей. Учитывается возможность появления большого числа возможных ошибок распознавания символов. Требуется найти границы максимального числа полей, одновременно минимизируется ошибка нахождения границ каждого поля. На втором проходе за счет параметризации распознавания достигается улучшение качества распознавания полей в найденных границах. В качестве параметров могут выступать алфавит распознавания или параметры обработки изображения в локальной области.

1. Описание документов с помощью структурных элементов и отношений

В работах [5, 6] предлагается подход к привязке полей, основанный на описании докумен-

та с помощью атрибутов его структурных элементов (часто употребляемые в печатных документах текстовые и графические объекты) и отношений между структурными элементами (СЭ). Рассмотрим кратко предложенный авторами [5, 6] поход к привязке полей документа. Атрибутами текстовых СЭ являются:

– 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.

Модель документа включает в себя:

– перечисление СЭ и их атрибутов, в том числе о расположении СЭ на странице;

– относительное расположение СЭ на листе.

Результатом работы алгоритма привязки является однозначное сопоставление каждого СЭ с конкретными геометрическими координатами на листе бумаги [5]. Рассматривались типы отношений между СЭ:

1. «Правее», «левее», «выше», «ниже»;

2. «Справа», «слева», «над», «под»;

3. «Соединены» для линий разграфки;

4. «До» («слева» либо «выше»), «после» («справа» либо «ниже»).

Совокупность отношений определяет граф отношений СЭ документа. В работе [5] граф отношений строился вручную. С помощью графа проводится поиск кандидатов, сопоставленных с очередным СЭ. Порядок выбора элементов был несущественен.

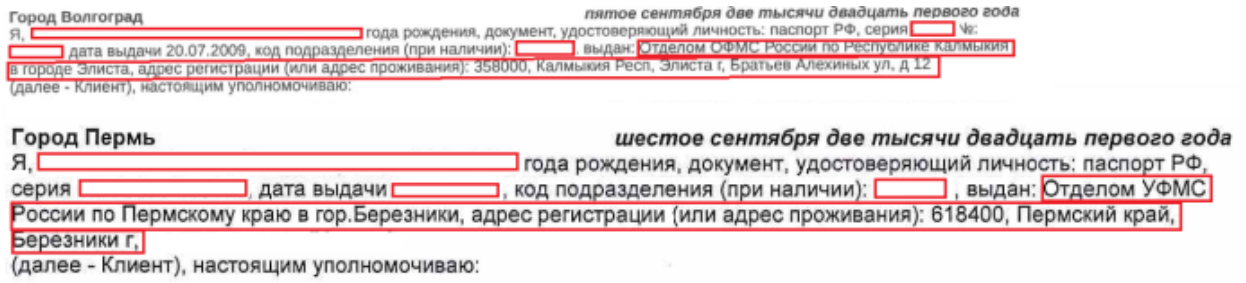
Алгоритм привязки каждого СЭ включает поиск по атрибутам и поиск по отношениям. Поиск по атрибутам генерирует начальное множество альтернативных местонахождений СЭ с оценками. Поиск по отношениям уточняет оценку каждой альтернативы в зависимости от степени удовлетворения отношениям, в которых СЭ участвует. Приведем несколько примеров описаний СЭ из [5]:

```
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;
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;
ПОЛЕ поле_терр_орган;
ПОЛЕ поле_рег_номер;
ПОЛЕ поле_внж_серия СИНТАКСИС=[0-9][0-9];
ОТНОШЕНИЕ линия1 ПОД заголовок
РАССТ_ОТН_МАКС = 0.1;
ОТНОШЕНИЕ регистрационный ПОД линия2_1
РАССТ_ОТН_МАКС = 0.01;
```

Особенностью реализованного [5] является ее способность работать с документами, структурные элементы которых не имеют фиксированного расположения на бумаге, а также с документами, примеры которых не представлены. Недостаток описанного подхода состоит в невозможности привязки гибких документов, отличающихся от исходного шаблона, как из-за существенного изменения атрибутов СЭ, так и их взаимного расположения (примеры на Рис. 1). Также авторы не рассматривают проблемы привязки статических текстовых СЭ к словам, распознанных с большим числом ошибок.

В данной работе предлагается другой способ привязки полей документа, также основанный на описаниях структурных элементов и отношениях между ними. Основное отличие предложенной модели документа по сравнению с [5] состоит в возможности группировки СЭ и в указании дополнительных атрибутов СЭ для успешного сопоставления текстового СЭ и слов, распознанных с ошибками.

Модель делового документа является иерархической. Верхним уровнем является *фрагмент* – часть страницы документа, отделенная от других зон линиями разграфки или проме-


Город Волгоград пятое сентября две тысячи двадцать первого года
Я, года рождения, документ, удостоверяющий личность: паспорт РФ, серия №:
дата выдачи 20.07.2009, код подразделения (при наличии): , выдан: Отделом ОФМС России по Республике Калмыкия
в городе Элиста, адрес регистрации (или адрес проживания): 358000, Калмыкия Респ, Элиста г, Братьев Алехиных ул, д 12
(далее - Клиент), настоящим уполномочиваю:

Город Пермь шестое сентября две тысячи двадцать первого года
Я, года рождения, документ, удостоверяющий личность: паспорт РФ,
серия , дата выдачи , код подразделения (при наличии): , выдан: Отделом УФМС
России по Пермскому краю в гор.Березники, адрес регистрации (или адрес проживания): 618400, Пермский край,
Березники г,
(далее - Клиент), настоящим уполномочиваю:

Рис. 1. Примеры группы СЭ в виде многострочного параграфа

жутками. Как правило, фрагменты являются частью дизайна документа и предназначены для группировки нескольких СЭ. Фрагменты содержат *составные объекты* и *локальные* СЭ, которые не принадлежат ни одному из составных объектов. К составным объектам относятся текстовые строки и многострочные объекты (*параграфы*). Параграфы могут со держать как известное заранее, так и переменное число строк. Локальный СЭ отличается от соседних элементов в некоторой окрестности, например, локальным СЭ является штрих-код. Строки и параграфы состоят из двух типов СЭ: текстовых (слов) и графических (линии разграфки). Возможны строки, не содержащие слов и состоящие только из графических СЭ.

Для описания документа применяется некоторый формальный язык привязки. Предложенный в данной работе язык привязки применятся в системе SDK Smart Document Engine [7] для настройки распознавания гибких документов. В языке СЭ называются объектами или составными объектами, а описания – дескрипторами.

Основные объекты относятся к следующим типам:

– терм (model_anchor – уникальное для фрагмента слово, model_similar – уникальное для строки или параграфа слово, model_twins – неуникальное для строки или параграфа слово, model_mandatory – обязательное для строки или параграфа слово);
– словарное слово (model_dict);
– недопустимое для составного объекта слово (model_disable);
– поле (model_field);
– линия разграфки (model_line);
– цепь (определение *цепи* будет определено в разделе 1, model_chain);
– параграф (model_paragraph);
– параграф (model_string);
– фрагмент (model_zone).

Каждый из объектов описывается как набор атрибутов:

– тип объекта;
– идентификатор объекта, уникальный для данного типа;
– слово (может отсутствовать для линий разграфки и составных объектов);
– рамка объекта;
– имя словаря или группы слов;
– ссылки на другие объекты.

Каждому объекту могут быть приписаны один или несколько тэгов (model_tag), содержащие произвольные атрибуты или параметры основного объекта.

В процессе привязки дескрипторы объектов сопоставляются с объектами, извлеченными из образа оцифрованной страницы документа, такими как:

– распознанное слово – последовательность символов с оценками надежности распознавания и рамкой каждого символа и слова в целом;
– найденный отрезок – рамка и атрибуты отрезка;
– найденная пометка – рамка пометки;
– найденный фрагмент – рамка фрагмента.

2. Дескрипторы строк и параграфов

Для жестких форм (документов с фиксированной геометрией) для задач привязки используются локальные особенности, задаваемые в виде особых точек различного типа (SURF, YAPE, YOLO, SIFT, ASIFT). Дескрипторы особых точек могут быть достаточно сложными для детектирования [8, 9]. Границы полей привязываются с по-

мощью набора особых точек, удовлетворяющих набору отношений. Предполагается, что в жесткой форме после оцифровки документа сохраняются отношения между парами точек. Отношения между парами точек в гибком документе вообще говоря не сохраняются. Это вытекает из возможности менять дизайн документов. Основными причинами изменения структуры гибкого документа являются:

– изменения характеристик шрифтов;

– изменения межстрочного расстояния, полей колонок текста;

– замены слов статического текста;

– изменения взаимного расположения слов текста и линий разграфки.

Аналогично особым точкам для жестких форм для привязки гибких документов возможно применение текстовых особых точек [10, 11]. Пара {T(W), B(W)}, в которой T(W) – слово, являющееся дескриптором текстовой особой точки, а B(W) – рамка, состоящая из координат четырехугольника, ограничивающего образ текстовой особой точки W, определяет дескриптор особой точки W. Детектором текстовой особой точки является процедура распознавания с помощью OCR. Сравнение двух текстовых особых точек проводится с помощью модифицированного расстояния Левенштейна. Предложенные модификации учитывают возможность сравнения текстовой особой точки со словами из распознанных зашумленных образов документов с большим числом ошибок распознавания. Предложенное в [11] модифицированное расстояние Левенштейна эффективно различает слова, близкие по классической метрике Левенштейна.

Некоторые слова, отличающихся от соседних слов в некоторой окрестности (документ, фрагмент, параграф, строка) являются уникальными при сравнении с помощью выбранной метрики. Разумеется, существуют неуникальные слова, повторяющиеся в некоторой окрестности. Для привязки полей, ограниченных неуникальными словами, необходимо использовать составные дескрипторы (*цепи*). Цепь C состоит из последовательности *термов*

$$C = \{Tm_1(C), Tm_2(C), \dots Tm_n(C)\}.$$

Терм $Tm_i(C)$ определяется как множество дескрипторов текстовых особых точек $W_1(Tm_i(C))$, $W_2(Tm_i(C))$,… и нескольких порогов $d_1(Tm_i(C))$, $d_2(Tm_i(C))$ … . Пороги используются при сравнении взаимного расположения данного терма Tm_i и предыдущего терма Tm_{i-1} при $i>1$. Задание отношений между термами не является обязательным. Каждый из указанных порогов $d_k(Tm_i(C))$ является параметром условия

$$\rho^k_T(Tm_{i-1}(C),Tm_i(C))<d_k(Tm_i(C)) \qquad (1)$$

где ρ^k_T – одна из возможных метрик для вычисления расстояния между двумя термами. Примерами таких метрик являются: количество слов между двумя термами или расстояние, вычисляемое с помощью рамок текстовых особых точек, входящих в термы. Термы могут обладать атрибутами для различения от других термов при сравнении. *Привязка* терма определяется как выбор слов распознанного документа, сходных с термом. Привязка проводится с помощью расстояния Левенштейна в сочетании с набором отношений с другими термами. Термам соответствуют объекты с типами model_anchor, model_similar, model_twins, model_mandatory, model_disable. Атрибутами термов являются:

– одно или несколько слов, использованных в шаблоне документа;

– рамки терма, координаты которых могут использоваться при сравнении;

– метрики ρ^k_T, и пороги сравнения (1), задаваемые с помощью объекта model_tag.

Цепи соответствуют объектам с типами model_chain, атрибутами цепей являются:

– рамка цепи, ограничивающая область поиска цепи;

– множество термов.

Простейшая цепь состоит из единственного обобщенного терма. Определение цепи позволяет описать уникальную последовательность термов, отличающуюся от других цепей в контексте документа или части документа (строк, параграфов, фрагментов).

В процессе привязки цепи C и распознанного документа участвуют дескриптор цепи и распознанные слова. Кандидаты W^{REC}_q на роль терма $Tm_i(C)$ проверяются на соответствие последовательности термов в цепи C, а именно из всех кандидатов на роль терма выбираются такие, что оценка последовательности термов цепи C минимизируется:

$$\delta(C) = \max(\rho_{LEV}(T(Tm_i(C)), W^{REC}{}_q)) \rightarrow \min. \quad (2)$$

При вычислении оценки привязки цепи в форме (2) требуется привязка всех термов цепи, то есть в цепи все термы должны иметь тип model_mandatory. В (2) каждый терм вносит штраф $\rho_{LEV}(T(Tm_i(C)), W^{REC}{}_q)$ в случае несоответствия к привязанному распознанному слову. В случае идеальной привязки терма штраф равен нулю. Для сильно зашумленных или сильно искаженных образов документов это сделать невозможно из-за отсутствия привязки некоторых термов. Часть непривязанных термов с типом отличным от model_mandatory и model_disable игнорируются при расчете более слабой оценки:

$$\delta_2(C_p) = \Sigma(\rho_{LEV}(T(Tm_i(L'_i, C_p)), W^{REC}{}_q) \geq d(Tm_i(C))) \rightarrow \min.$$

Если же был привязан терм $Tm_i(C_p)$ с типом model_disable, то $\delta_2(C_p) = 0$, то цепь C_p считается непривязанной.

Однострочное или многострочное поле F описывается дескриптором поля $\{C_1; F; C_2\}$. Левая граница поля F определяется с помощью рамок самого правого терма цепи C_1, правая граница – самого левого терма цепи C_2. Цепи C_1 и C_2 являются *опорными* элементами для поля F. Если граничные термы не привязаны, для привязки могут быть взяты другие термы цепей C_1 и C_2. В таком случае может увеличиваться *ошибка* привязки поля (различие между прогнозной рамкой и границами реального поля в образе документа). В дескрипторе поля может отсутствовать одна из цепей C_1 или C_2. В таких случаях граница, соответствующая полю, задается границей зоны или документа. Верхняя и нижняя граница одной из строк поля F задаются границами текстовых строк, входящих в цепи или размещенных между нижней границей цепи C_1 и верхней границы цепи C_2. Для строки или параграфа, содержащих несколько полей дескриптор выглядит следующим образом:

$$D_{TEXT}(F_1, F_2, \ldots, F_{p2}) = \begin{cases} C_1^{top}; C_2^{top}; \ldots C_{p1}^{top}; \\ C_1; F_1; C_2; F_2; \ldots C_{p2}; F_{p2}; C_{p2+1}; \\ C_1^{bottom}; C_2^{bottom}; \ldots C_{p3}^{bottom}. \end{cases} \quad (3)$$

В дескрипторе (3) цепи $C^{top}{}_1$, $C^{top}{}_2$,… $C^{top}{}_{p1}$ служат для ограничения сверху области поиска опорных цепей C_1, C_2, … C_{p2+1}. Аналогично, цепи $C^{bottom}{}_1$, $C^{bottom}{}_2$, … $C^{bottom}{}_{p3}$ ограничивают поиск опорных цепей снизу. Оценка привязки дескриптора D_{TEXT} основана на оценках привязанных цепей этой группы

$$\delta(C_1, C_2, \ldots, C_{p2}) = \min(\delta(C_k)).$$

Дескрипторы (3) соответствуют объектам с типами model_string, model_paragraph, поля F_1; F_2; … F_{p2} имеют тип model_field. Атрибутами дескриптора (3) являются:

– рамка дескриптора, ограничивающая область поиска цепи;

– множество цепей для ограничения области поиска сверху;

– множество цепей для ограничения области поиска снизу;

– множество опорных цепей;

– множество полей.

Эффективность привязки поля определяется применением уникальных цепей и применением функции сравнения ядра терма и распознанного слова, учитывающие возможные ошибки распознавания.

3. Дескрипторы отрезков

В случае плохого распознания символов для поиска полей строки, привязку полей может улучшить применение отрезков линий (отрезков). Поиск объектов документов, ограниченных линиями, в рассмотренных ниже случаях упрощает привязку полей, ячеек таблицы или форм более сложной формы. Атрибутами отрезка S являются:

– ориентация (вертикальный, горизонтальный);

– тип (сплошной, пунктирный, состоящий из точек);

– рамка отрезка;

– длина и толщина отрезка;

– ориентировочное размещение в образе документа или зоны.

Известны алгоритмы извлечения отрезков из образа документа [12 - 14]. Также известны алгоритмы, формирующие таблицу или сложную форму из массива выделенных отрезков линий.

Чаще всего один единственный отрезок не отличается от множества сходных с ним отрезков и не может быть рассмотрен в качестве опорного элемента. Рассмотрим объекты, состоящие из текстовых объектов и отрезков. На Рис. 2. приведены примеры фрагментов универсального передаточного документа. В этих примерах привязка полей с применением цепей затруднительна из-за потерь большого числа символов. Это объясняется существенным зашумлением образа, связанным не с оцифровкой, а с образом печати. Для таких случаев привязку полей улучшает использование комбинированных объектов, состоящих из цепей и отрезков. Комбинированным объектом является следующая совокупность цепей и отрезков:

$$D(F_1, F_2, \dots, F_{p2}) = \begin{cases} C_1^{top};\ C_2^{top};\dots C_{p1}^{top}; \\ C_1;\frac{F_1}{S_1};\ C_2;\frac{F_2}{S_2};\ \dots\ C_{p2};\frac{F_{p2}}{S_{p2}};\ C_{p2+1}; \\ C_1^{bottom};\ C_2^{bottom};\dots C_{p3}^{bottom}. \end{cases} \quad (4)$$

В описании (4) отрезки S_1, S_2, … S_{p2} должны являться частью одной линии S. Расстояние до линии S каждой из рамок термов цепей C_1, C_2, … C_{p2+1} ограничено. Элементы (4) выбираются и настраиваются при обучении таким образом, чтобы дескриптор был локально особенным. Некоторые из цепей и отрезков (4) могут быть необязательными при привязке. Оценка привязки базируется на оценках привязки каждой из цепей и оценки привязки набора отрезков. Набор отрезков оценивается соответствием числа и отношений размеров отрезков дескриптора комбинированного дескриптора и извлеченных из образа документа отрезков. Оценка привязки дескриптора нескольких полей основана на оценках привязанных цепей этой группы и оценке привязи совокупности отрезков

$$\delta(C_1, C_2, \dots, C_{p2}) = \min(\delta(C_k)) - \delta(S_1, S_2, \dots, S_{p2+1}).$$

Дескрипторы (3) соответствуют объектам с типами model_string, model_paragraph, атрибутами дескриптора (3) являются:

– рамка дескриптора, ограничивающая область поиска цепи;

– множество цепей для ограничения области поиска сверху;

– множество цепей для ограничения области поиска снизу;

– множество опорных цепей;

– множество отрезков;

– множество полей, размещенных над отрезками.

Рассмотрим случай упрощенного представления дескриптора (4):

$$D(F_1, F_2, \dots, F_{p2}) = \begin{cases} C_1^{top};\ C_2^{top};\dots C_{p1}^{top}; \\ \frac{F_1}{S_1};\frac{F_2}{S_2};\ \dots\ \frac{F_{p2}}{S_{p2}}; \\ C_1^{bottom};\ C_2^{bottom};\dots C_{p3}^{bottom}. \end{cases} \quad (5)$$

В дескрипторе (4) опорными элементами являются верхние и нижние цепи и последовательность отрезков S_1, S_2, … , S_{p2+1}. Предполагается, что проекции на вертикальную ось всех отрезков одинаковы или незначительно различаются. В процессе привязки участвуют извлеченные отрезки-кандидаты s_1, s_2, … s_r, размещенные в области, ограниченной цепями $C^{top}{}_1$, $C^{top}{}_2$, … $C^{top}{}_{p1+1}$ и $C^{bottom}{}_1$, $C^{bottom}{}_2$, … $C^{bottom}{}_{p3+1}$. Координаты отрезков-кандидатов нормализу-

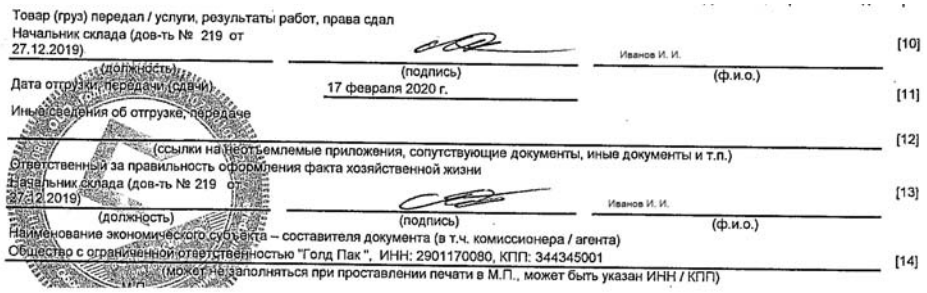

Товар (груз) передал / услуги, результаты работ, права сдал
Начальник склада (дов-ть № 219 от 27.12.2019) [10]
(должность) (подпись) Иванов И. И. (ф.и.о.)
Дата отгрузки, передачи (сдачи) 17 февраля 2020 г. [11]
Иные сведения об отгрузке, передаче [12]
(ссылки на неотъемлемые приложения, сопутствующие документы, иные документы и т.п.)
Ответственный за правильность оформления факта хозяйственной жизни
Начальник склада (дов-ть № 219 от 27.12.2019) [13]
(должность) (подпись) Иванов И. И. (ф.и.о.)
Наименование экономического субъекта – составителя документа (в т.ч. комиссионера / агента)
Общество с ограниченной ответственностью "Голд Пак ", ИНН: 2901170080, КПП: 344345001 [14]
(может не заполняться при проставлении печати в М.П., может быть указан ИНН / КПП)

Рис. 2. Фрагмент универсального передаточного документа

ются на ширину фрагмента или на ширину страницы документа. Проводится кластер-анализ отрезков-кандидатов $\{s_1, s_2, \dots\}$. Группировка проводится с помощью близости проекций отрезков-кандидатов на вертикальную ось. Далее проводится оценка соответствия каждого из получившихся кластеров отрезков с $Cl_t=\{s^t_1, s^t_2, \dots\}$ и отрезков дескриптора $S_1, S_2, \dots, S_{p2+1}$. Если количество отрезков в кластере совпало с количеством отрезков дескриптора, то оценка привязки кластера вычисляется как сумма невязок длин пар отрезков:

$$\delta(Cl_t, D(F_1, F_2, \dots, F_{p2+1})) = \sum_j \left| h(S_j) - h(s_j) \right| \quad (6)$$

Также рассматриваются случаи, когда одному отрезку S_j соответствуют несколько отрезков-кандидатов $s^t_z, s^t_{z+1}, \dots, s^t_{z+l(j)}$, и случаи, когда одному отрезку-кандидату s_z соответствуют несколько от резков $S_v, S_{V+1}, \dots, S_{V+r(z)}$. Для этих случаев рассматриваются объединения нескольких отрезков или отрезков-кандидатов таким образом. Комбинирование основано на переборном алгоритме над отрезками дескриптора и отрезками-кандидатами. Для случая инвариантности размеров отрезков (с точностью до масштабирования) при печати документов с помощью минимизации штрафа (6) выбираются наборы отрезков-кандидатов Cl_t, границы которых наилучшим образом совмещаются с границами $S_1, S_2, \dots$.

Для документов, подобных показанным на Рис. 2, привязка с помощью описанного комбинированного дескриптора (3) улучшается по отношению к привязке с использованием текстового дескриптора (2) по следующим причинам:

– образы массивных отрезков на фоне подписей и печатей извлекаются более надежно, чем образы символов и слов;

– использование нескольких цепей для ограничения области поиска набора отрезков надежно ограничивают область поиска отрезков.

Пример группы полей, которые можно описать комбинированным дескриптором приведен на Рис. 3. Дескриптор для этого составного объекта, включающего два ключевых слова и четыре отрезка, записывается как

$$D(F_{295}, F_{300}, F_{296}, F_{306},) = \begin{cases} \{VZ_{0397}; VZ_{0314}\}; \\ \frac{F_{295}}{S_{295}}; \frac{F_{300}}{S_{300}}; \frac{F_{296}}{S_{296}}; \frac{F_{301}}{S_{301}}. \end{cases}$$

Успех привязки с помощью дескриптора (4) и (5) обеспечен в случаях, когда в условиях зашумления образа документа метод извлечения отрезков надежно находит отрезки. Применение дескрипторов с пунктирными типами отрезков для зашумленных изображений может оказаться менее эффективным.

4. Дескрипторы пометок

Пометка является одним из самых используемых графических элементов при оформлении документа. Область пометки ограничена четырьмя отрезками. Отрезки пометки могут быть как сплошными, так и пунктирными.

Различие пустой или заполненной области пометки проводится с помощью методов выделения контуров [15], CNN [16] или метода Виола и Джонса [17, 18]. Будем считать, что на первом этапе обработки документа наряду с распознаванием слов и детектированием символов были найдены рамки пометки. При детектировании возможны ошибки как пропуска пометки, так и ложного детектирования. При избыточном детектировании кандидатов пометки также используются комбинированные дескрипторы.

В дескрипторе (3) каждое из полей F может интерпретироваться как прогнозная рамка пометки. Пусть дескриптор (3) был привязан. Привязка содержащихся в нем пометок состоит в следующем. Для прогнозной рамки привязанного поля F выберем пометку-кандидат, площадь пересечения которого с рамкой F является наибольшей. Рамка пометки может искажаться

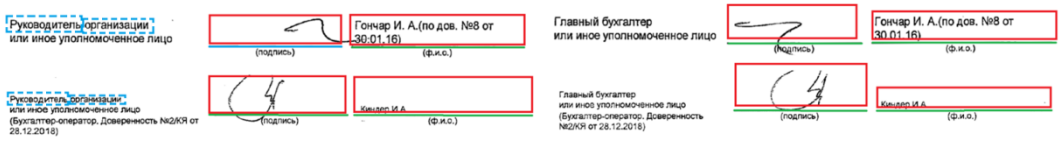


Рис. 3. Пример группы полей, заданных комбинированным дескриптором

из-за рукописного заполнения. В этом случае для распознавания пометки используется прогнозная рамка.

При большом количестве ошибок распознавания символов и детектирования отрезков могут не быть привязаны цепи в дескрипторах (2), (3), (4). В некоторых случая, например, для группы вертикальных пометок могут быть применены дескрипторы следующего вида:

$$D_{\mathrm{CB}}(\mathrm{F}_1, \mathrm{F}_2, \dots, \mathrm{F}_{p2}) = \begin{cases} \mathrm{C}_1^{\mathrm{top}};\ \mathrm{C}_2^{\mathrm{top}};\ \dots\ \mathrm{C}_{p1}^{\mathrm{top}}; \\ \mathrm{C}_1; \mathrm{F}_1; \mathrm{C}'_1; \\ \mathrm{C}_2; \mathrm{F}_2; \mathrm{C}'_2; \\ \dots \\ \mathrm{C}_{p2+1}; \mathrm{F}_{p2};\ \mathrm{C}'_{p2+1}; \\ \mathrm{C}_1^{\mathrm{bottom}};\ \mathrm{C}_2^{\mathrm{bottom}};\ \dots\ \mathrm{C}_{p3}^{\mathrm{bottom}}. \end{cases} \quad (7)$$

В дескрипторе (7) основными элементами являются верхняя и нижняя ограничивающие цепи C^{top}_1, C^{top}_2, … C^{top}_{p1+1} и C^{bottom}_1, C^{bottom}_2, … C^{bottom}_{p3+1} и группа вертикальным образом расположенных чек-боксов. Пример такой группы пометок показан на Рис. 4. Цепи C_1,C'_1, C_2,C'_2, … C_{p2+1},C'_{p2+1} не являются обязательными элементами при привязке группы пометок. Однако привязка какой-либо из цепей C_1,C'_1, C_2,C'_2, … C_{p2+1},C'_{p2+1} упрощает привязку в случае, когда не все пометки из группы были детектированы.

5. Дескрипторы фрагмента и страницы

Предложенный способ позволяет создавать дескрипторы для классификации страницы или фрагмента страницы. Иерархическое разбиение страницы делового документа на фрагменты – области, которые отделены друг от друга разделяющими линиями или большими промежутками, упрощает применение переборных алгоритмов. Основной причиной является уменьшение времени перебора из-за существенно меньшего количество слов в каждом из фрагментов по сравнением с количеством слов на странице.

В предположении, что известно некоторое множество фрагментов-кандидатов, являющихся подмножествами слов страницы документа, зададим дескриптор фрагмента Z с типом model_zone следующим образом:

$$\mathrm{D(Z)} = \{C_1(\mathrm{Z}), C_2(\mathrm{Z}), \dots\},$$

где $C_i(\mathrm{Z})$ – цепи, определенные в разделе 1. Привязка фрагмента Z проводится в случае, когда привязана хотя бы одна из цепей $C_1(\mathrm{Z})$, $C_2(\mathrm{Z})$,…. Оценка привязки фрагмента Z задается как минимальная из оценок привязки цепей $C_1(\mathrm{Z})$, $C_2(\mathrm{Z})$,…. Возможен случай, когда будет привязано более одного фрагмента-кандидата. Для разрешения таких конфликтов применяются термы с типом model_disable.

Классификация типа страницы, то есть выбор известного типа страницы документа, проводится аналогично. Для каждого из заданных заранее дескрипторов страницы

$$\mathrm{D_1(P)} = \{C_{11}(\mathrm{P}), C_{12}(\mathrm{P}), \dots\}$$
$$\mathrm{D_2(P)} = \{C_{21}(\mathrm{P}), C_{22}(\mathrm{P}), \dots\}$$
$$\dots$$

проводится привязка к распознанной странице и вычисляются оценки соответствия каждому типу документа. Оценка соответствия типу k задается как минимальная из оценок привязки цепей $C_{k1}(\mathrm{Z})$, $C_{k2}(\mathrm{Z})$,…. Далее оценки упорядочиваются и выбирается минимальная из оценок, соответствующих наиболее близкому классу или несколько наименьших оценок для многоклассовой классификации.

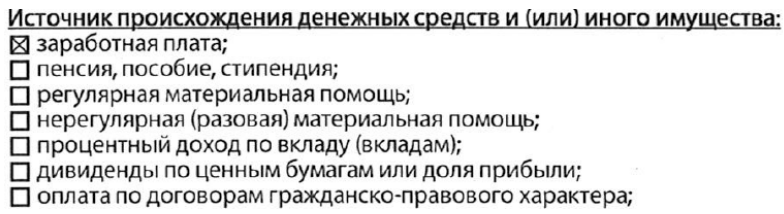

Источник происхождения денежных средств и (или) иного имущества:
☒ заработная плата;
☐ пенсия, пособие, стипендия;
☐ регулярная материальная помощь;
☐ нерегулярная (разовая) материальная помощь;
☐ процентный доход по вкладу (вкладам);
☐ дивиденды по ценным бумагам или доля прибыли;
☐ оплата по договорам гражданско-правового характера;

Рис. 4. Примеры вертикальной группы пометок

Табл. 1. Точность привязки полей тестового набора универсальных передаточных документов

Документ	Количество документов	Число полей	Точность распознавания полей	Точность привязки полей
с использованием отрезков	1545	55648	66,77 %	82,53 %
без использования отрезков	1545	55648	49,15 %	66,26 %

6. Результаты экспериментов

Предложенный метод анализа структуры документов был протестирован на собственном тестовом наборе, состоящем из образов документов типа «универсальный передаточный документ», отсканированных с оптической плотностью от 100 до 300 dpi с различным качеством оцифровки. Исследовались точность распознавания и точность привязки полей. Рассматривались случаи с использованием отрезков в качестве элементов комбинированных дескрипторов и без использования отрезков. Для распознавания использовался SDK Smart Document Engine [7]. Полученные результаты сведены в Табл. 1, иллюстрирующую эффективность предложенного метода.

Точность привязки фрагментов, подсчитанная на том же самом приватном датасете составила 99,5%. Точность привязки пометок, подсчитанная на приватном датасете документов типа «KYC» (анкета для идентификации клиента) составила 99,2%. Разумеется, точность детектирования пометок, равная 97,3%, существенно ниже точности привязки пометок.

Заключение

Предложенный метод анализа структуры документов применим для распознавания сильно зашумленных и искаженных образов документов. В частности, метод позволяет детектировать границы полей в черно-белых копиях цветных документов. Использование нескольких типов структурных элементов для привязки обеспечивает эффективность поиска и прогноза границ однострочных и многострочных полей. Эффективность достигается за счет использования комбинированных дескрипторов. Комбинированные дескрипторы основаны на простых дескрипторах слов статического текста, отрезков линий разграфки, чек-боксов.

Предложенный метод был опробован на деловых документах нескольких типов, таких как

- налоговые декларации;
- финансовые отчеты;
- торговые документы;
- банковские документы.

Точность предложенного метода была оценена на собственных датасетах. Точность привязки при использовании комбинированных дескрипторов находится в диапазоне от 80 до 99% для различных полей. Наихудшие результаты соответствуют сложным случаям зашумления печатями и подписями образов слов черно-белой копии цветного делового документа.

Литература

1. Башкатова, А. Цифровая экономика плодит все больше бумаг: Россияне не скоро перестанут носить в организации справки / Независимая Газета. – 2019 – 14 ноя. https://www.ng.ru/economics/2019-11-14/4_7727_paper.html (Доступ 22.09.2022)
2. Rusiñol M., Frinken V., Karatzas, D., Bagdanov, A. D., Lladós, J.: Multimodal page classification inadministrative document image streams. In: IJDAR. Vol. 17(4), pp. 331 Image Classification by Mixed Finite Element Method and Orthogonal Legendre Moments 341. (2014). https://doi.org/10.1007/s10032-014-0225-8
3. Jain, R., Wigington, C.: Multimodal Document Image Classification. pp. 71–77. (2019). https://doi.org/10.1109/ICDAR.2019.00021
4. Qasim, S. Rukh., Mahmood, H., Shafait, F.: Rethinking Table Recognition using Graph Neural Networks. pp. 142–147. (2019). https://doi.org/10.1109/ICDAR.2019.00031
5. Марченко, А. Е., Ершов, Е. И., Гладилин, С. А. Система разбора документов, заданных атрибутами элементов структуры и отношениями между элементами / Труды ИСА РАН, Т. 67(4). сс. 87–97. (2017).
6. Postnikov V. V.: Identification and Recognition of Documents with a Predefined Structure // Pattern Recognition and Image Analysis. Vol. 13. № 2. pp. 332–334. (2003)
7. Smart Document Engine – automatic analysis and data extraction from business documents for desktop, server and mobile platforms / https://smartengines.com/ocr-engines/document-scanner (Доступ 22.09.2022)

8. Bellavia, F.: SIFT Matching by Context Exposed. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022). https://doi.org/10.1109/TPAMI.2022.3161853
9. Bay, H., Tuytelaars, T., Van Gool, Luc.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding - CVIU. Vol. 110. No. 3, pp. 404–417. (2006).
10. Slavin, O., Andreeva, E., Paramonov, N.: Matching Digital Copies of Documents Based on OCR, 2019 XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP), pp. 177–181, (2019). https://doi.org/10.1109/CSCMP45713.2019.8976570
11. Slavin, O., Arlazarov, V., Tarkhanov, I.: Models and Methods Flexible Documents Matching Based on the Recognized Words. Cyber-Physical Systems: Advances in Design & Modelling. Springer Nature Switzerland AG. Vol. 350, pp. 173–184 (2021). https://doi.org/10.1007/978-3-030-67892-0_15
12. Matas, J., Galambos, C., Kittler, J.: Robust Detection of Lines Using the Progressive Probabilistic Hough Transform, Computer Vision and Image Understanding, Vol. 78, Issue 1, pp. 119–137, (2000). https://doi.org/10.1006/cviu.1999.0831
13. Grompone von Gioi, R., Jakubowicz, J., Morel, JM. et al.: On Straight Line Segment Detection. J Math Imaging Vis. Vol. 32, pp. 313–347. (2008). https://doi.org/10.1007/s10851-008-0102-5
14. Grompone von Gioi R., Jakubowicz J., Morel J.-M., Randall G.: LSD: A Fast Line Segment Detector with a False Detection Control / IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 32, Issue 4. pp. 722–732. (2010). https://doi.org/10.1109/TPAMI.2008.300
15. Emaletdinova, L. & Nazarov, M.: Construction of a Fuzzy Model for Contour Selection. Construction of a Fuzzy Model for Contour Selection. In: Kravets, A.G., Bolshakov, A.A., Shcherbakov, M. (eds) Cyber-Physical Systems: Intelligent Models and Algorithms. Studies in Systems, Decision and Control, Vol. 417. pp. 243–246. (2022). https://doi.org/10.1007/978-3-030-95116-0_20.
16. Zlobin, P., Chernyshova, Y., Sheshkus A., Arlazarov V. V.: Character sequence prediction method for training data creation in the task of text recognition. Proc. SPIE 12084, Fourteenth International Conference on Machine Vision (ICMV 2021), 120840R. (2022). https://doi.org/10.1117/12.2623773
17. Matalov, D., Usilin, S., Arlazarov, V.V.: About Viola-Jones image classifier structure in the problem of stamp detection in document images. Proc. SPIE 11605, Thirteenth International Conference on Machine Vision, 116050V (2021). https://doi.org/10.1117/12.2586842
18. Arlazarov, V., Voysyat, Ju. S., Matalov, D., Nikolaev, D., Usilin, S.A.: Evolution of the Viola-Jones Object Detection Method: A Survey. Vol. 14. pp. 52–23. (2021). https://doi.org/10.14529/mmp210401

Славин Олег Анатольевич. Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", г. Москва, Россия. Главный научный сотрудник, доктор технических наук. Количество печатных работ: 77 (в т.ч. 1 монография). Область научных интересов: распознавание образов, информационные системы. E-mail: oslavin@isa.ru

Object Descriptors for Linking Structural Elements of Noisy Document Images

O. A. Slavin [I,II]

[I] Federal State Institution "Federal Research Center" Informatics and Management "of the Russian Academy of Sciences", Moscow, Russia
[II] "Smart Engines", Moscow, Russia

Abstract. The problem of extracting filling elements (fields) from a recognized image of a document with the help of descriptors - descriptions of one or more structural elements is considered. Structural elements can be words of static text and scribble lines used to shape the design of a document. Business documents with a simplified structure and a limited vocabulary are considered. Flexible business documents that allow significant modifications to the page design are considered. Descriptors are created taking into account a significant number of possible errors in document page recognition. Combined descriptors consisting of several terms and line segments are described. A binding algorithm based on descriptors is given. It is experimentally shown that the extraction of combined descriptors improves the accuracy of recognition of document fields during recognition by 17%, and the accuracy of extracting information from the document image by 16%. The SDK Smart Document Engine was used as OCR in the experiment.

Keywords: virtual reality, augmented reality, virtual reality helmet, immersiveness, virtual object, heptic technologies, content.

DOI 10.14357/20718632220402

References

1. Bashkatova, A. Cifrovaya ekonomika plodit vse bol'she bumag: Rossiyane ne skoro perestanut nosit' v organizacii spravki // Nezavisimaya Gazeta. – 2019 – 14 ноя. https://www.ng.ru/economics/2019-11-14/4_7727_paper.html (accessed September 22, 2022).
2. Rusiñol M., Frinken V., Karatzas, D., Bagdanov, A. D., Lladós, J.: Multimodal page classification inadministrative document image streams. In: IJDAR. Vol. 17(4), pp. 331 Image Classification by Mixed Finite Element Method and Orthogonal Legendre Moments 341. (2014). https://doi.org/10.1007/s10032-014-0225-8.
3. Jain, R., Wigington, C.: Multimodal Document Image Classification. pp. 71–77. (2019). https://doi.org/10.1109/ICDAR.2019.00021.
4. Qasim, S. Rukh., Mahmood, H., Shafait, F.: Rethinking Table Recognition using Graph Neural Networks. pp. 142–147. (2019). https://doi.org/10.1109/ICDAR.2019.00031.
5. Marchenko A.E., Ershov E.I., Gladilin S.A. Sistema razbora dokumenta, zadannogo atributami strukturnykh elementov i otnosheniyami mezhdu strukturnymi elementami [The system for parsing a document specified by attributes of structural elements and the relations between structural elements] / Trudy ISA RAN, Vol 67, No 4, pp. 87-97. (2017).
6. Postnikov V. V.: Identification and Recognition of Documents with a Predefined Structure // Pattern Recognition and Image Analysis. Vol. 13. № 2. pp. 332–334. (2003).
7. Smart Document Engine – automatic analysis and data extraction from business documents for desktop, server and mobile platforms / https://smartengines.com/ocr-engines/document-scanner (accessed September 22, 2022).
8. Bellavia, F.: SIFT Matching by Context Exposed. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022). https://doi.org/10.1109/TPAMI.2022.3161853.
9. Bay, H., Tuytelaars, T., Van Gool, Luc.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding - CVIU. Vol. 110. No. 3, pp. 404–417. (2006).
10. Slavin, O., Andreeva, E., Paramonov, N.: Matching Digital Copies of Documents Based on OCR, 2019 XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP), pp. 177–181, (2019). https://doi.org/10.1109/CSCMP45713.2019.8976570.
11. Slavin, O., Arlazarov, V., Tarkhanov, I.: Models and Methods Flexible Documents Matching Based on the Recognized Words. Cyber-Physical Systems: Advances in Design & Modelling. Springer Nature Switzerland AG. Vol. 350, pp. 173–184 (2021). https://doi.org/10.1007/978-3-030-67892-0_15.
12. Matas, J., Galambos, C., Kittler, J.: Robust Detection of Lines Using the Progressive Probabilistic Hough Transform, Computer Vision and Image Understanding, Vol. 78, Issue 1, pp. 119–137, (2000). https://doi.org/10.1006/cviu.1999.0831.
13. Grompone von Gioi, R., Jakubowicz, J., Morel, JM. et al.: On Straight Line Segment Detection. J Math Imaging Vis. Vol. 32, pp. 313–347. (2008). https://doi.org/10.1007/s10851-008-0102-5.
14. Grompone von Gioi R., Jakubowicz J., Morel J.-M., Randall G.: LSD: A Fast Line Segment Detector with a False Detection Control / IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 32, Issue 4. pp. 722–732. (2010). https://doi.org/10.1109/TPAMI.2008.300.
15. Emaletdinova, L. & Nazarov, M.: Construction of a Fuzzy Model for Contour Selection. Construction of a Fuzzy Model for Contour Selection. In: Kravets, A.G., Bolshakov, A.A., Shcherbakov, M. (eds) Cyber-Physical Systems: Intelligent Models and Algorithms. Studies in Systems, Decision and Control, Vol. 417. pp. 243–246. (2022). https://doi.org/10.1007/978-3-030-95116-0_20.
16. Zlobin, P., Chernyshova, Y., Sheshkus A., Arlazarov V. V.: Character sequence prediction method for training data creation in the task of text recognition. Proc. SPIE 12084, Fourteenth International Conference on Machine Vision (ICMV 2021), 120840R. (2022). https://doi.org/10.1117/12.2623773.
17. Matalov, D., Usilin, S., Arlazarov, V.V.: About Viola-Jones image classifier structure in the problem of stamp detection in document images. Proc. SPIE 11605, Thirteenth International Conference on Machine Vision, 116050V (2021). https://doi.org/10.1117/12.2586842.
18. Arlazarov, V., Voysyat, Ju. S., Matalov, D., Nikolaev, D., Usilin, S.A.: Evolution of the Viola-Jones Object Detection Method: A Survey. Vol. 14. pp. 52–23. (2021). https://doi.org/10.14529/mmp210401.

Slavin O. A. Federal State Institution "Federal Research Center" Informatics and Management "of the Russian Academy of Sciences", Moscow, Russia. Chief Researcher, Doctor of Technical Sciences. Number of publications: 77 articles, 1 monograph. Research interests: pattern recognition, information systems. E-mail: oslavin@isa.ru