# Алгоритмы привязки полей при распознавании условно-жестких деловых документов

О. А. Славин

Федеральный исследовательский центр "Информатика и управление" РАН, Москва, Россия
ООО "Смарт Энджинс Сервис", Москва, Россия

**Аннотация.** Предложены определения гибких и жестких документов, используемые в технологиях ввода в компьютер деловых документов. Рассмотрены особенности создания, оцифровки и анализа жестких форм и жестких документов. Описаны границы применимости модели привязки изображений жестких документов, искаженных при оцифровке. Рассмотрена модель для привязки гибких документов, основанная на распознанных словах и графических примитивах, связанных набором отношений порядка. Классификация основана на различных способах подготовки деловых документов для печати. Описаны особенности привязки полей и распознавания для нескольких типов документов, таких как условно-жесткие документы, гибкие документы, продуцированные одной формой, гибкие документы, продуцированные малым и большим числом форм. Рассмотрен случай распознавания условно-жестких документов с применением технологий ввода гибких документов. Проведенные эксперименты показывают, что для некоторых полей пометок в условиях сильного зашумления и значительных искажений доля ошибок уменьшается в два раза.

**Ключевые слова:** распознавание документов, условно-жесткий документ, текстовая особая точка, пометка.

**DOI** 10.14357/20718632230404 **EDN** BXZTGE

## Введение

Задача распознавания документов по-прежнему является актуальной по причине роста объемов напечатанных на бумаге деловых документов и документов, удостоверяющих личность. Например, объемы потоков входящих и исходящих документов в крупных организациях могут достигать нескольких сотен страниц в день. В данной работе рассматриваются деловые документы, предназначенные для обмена данными с организациями и физическими личностями [1]. Мы будем определять документ как совокупность полей и статической информации. В работе рассматриваются деловые документы. Деловые документы характеризуются относительно простой структурой и ограниченным словарем статических текстов. Статическими элементами, прежде всего, являются слова статического текста. Статические слова группируются в строки, заголовки, абзацы и параграфы. Поля могут определяться как объект, который ограничен несколькими статическими элементами, такими как:

- слова статического текста;
- отрезки (линии подчеркивания);
- бар-коды,
- пометки (чек-боксы).

Извлечение информации из распознанных деловых документов имеет ряд особенностей:

- малый объем словаря слов статического текста;

- возможное значительное число ошибок распознавания;
- возможные ошибки детектирования графических элементов.

Постановка задачи распознавания страницы документа состоит в следующем. На основании распознавания текстовых объектов и найденных графических примитивов найти границы полей (областей заполнения) и извлечь информацию из областей полей. В работе рассматриваются технологии, основанные на описании документов [2]. Рассматриваются распознанные зашумленные изображения документов. Зашумлению подвержены образы символов, отрезки, пометки. Эта тематика в настоящее время является актуальной [3-6].

Распознавание образа делового документа может быть реализовано в виде следующих этапов:

- нормализация образа страницы, в том числе, поиск области документа и его приведение к прямоугольному виду;
- распознавание слов;
- извлечение графических объектов;
- классификация типа документа;
- поиск локальных особенностей;
- поиск границ полей документа известного типа с помощью границ локальных особенностей;
- извлечение или распознавание содержимого полей в найденных границах с помощью атрибутов полей;
- постобработка распознанных полей с помощью словарных моделей.

Критерием решения задачи распознавания является извлечение информации из границ максимального числа полей с наименьшим числом ошибок для каждого поля. Извлекаемая информация может иметь вид не только набора символов, но и границ найденного поля.

## 1. Жесткие и гибкие документы

Известно подразделение деловых документов на *жесткие* и *гибкие* [7]. В жестком документе оформление и статические тексты неизменны, а в гибком – допускаются различные изменения. Рассмотрим документы, созданные печатью информации на бланке, универсальном для всех документов определенного типа. Назовем *формой* бланк документа без заполнения. Жесткие формы могут быть напечатаны полиграфическим способом. К таким жестким документам относятся идентификационные документы, свидетельства, дипломы. Другим способом является использование при печати документа одного единственного шаблона, который не может изменяться пользователем, например, в формате PDF. Распечатка такого шаблона без заполнения является жесткой формой. Получение цифровой копии (*оцифровка*) при вводе в компьютер возможно как для документа, так и для формы. В процессе печати документов могут быть получены различные документы, соответствующих одной форме. Аналогично в процессе оцифровки могут быть получены различные изображения (экземпляры), соответствующие одному и тому же документу.

Рассмотрим два экземпляра $I_1$ и $I_2$ одного оцифрованного документа $I \in \mathbb{R}^2$ в предположении, что процесс оцифровки позволяет получать образы документов без потери частей документа. После оцифровки подвергнем изображение документа нормализации по форме и размеру. Различия в двух экземплярах одного документа определяются дефектами оцифровки и точностью нормализации образа документа. Рассмотрим отображение $f$ точек изображений $f$: $\mathbb{R}^2 \to \mathbb{R}^2$. Рассмотрим пары (A, B) из одного изображения $A \in I_1$, $B \in I_1$, таких что $d(A, B) > \delta$, где $d$ – метрика Евклида.

Если для соответствующих точкам (A, B) точек ($f$(A), $f$(B)) из другого изображения угол между векторами (A, B) и ($f$(A), $f$(B)) является малым и разница между расстояниями $d$(A, B) и $d(f(A), f(B))$ является малой:

$$\begin{cases} \arccos\left(\dfrac{\overrightarrow{AB}\cdot\overrightarrow{F(A)F(B)}}{|\overrightarrow{AB}|\cdot|\overrightarrow{F(A)F(B)}|}\right) < \theta, \\ \left| d(A,B) - d\big(f(A), f(B)\big)\right| < \varepsilon, \end{cases} \quad (1)$$

то отображение $f$ назовем *псевдоизометрическим* для рассматриваемого множества точек (A, B). В реальности из-за дефектов оцифровки и физических дефектов документов возможны сильные искажения изображений. Поэтому для далеких пар точек (A, B) возможны сильные изменения расстояний в их образах ($f$(A), $f$(B)). Параметры $\delta$, $\theta$ и $\varepsilon$ могут быть выбраны как для

всего изображения, так и для части. Минимальным значением ε является 1 (1 пиксель). Параметр δ может быть достаточно большим, например, расстояние между границами слов и строк в жесткой форме остается приблизительно одинаковым. Области изображений $I_1$ и $I_2$, для которых выполняется условие (1) с параметрами δ, θ и ε назовем *областью псевдоизометрии*.

*Жесткой формой* будем называть такую форму, для которой после оцифровки полученные экземпляры содержат области псевдоизометрии. Для этого определения нужно сделать несколько оговорок. Во-первых, мы предполагаем, что способ оцифровки, например, с помощью сканера или камеры мобильного устройства, позволяет получать «близкие» цифровые экземпляры формы. Во-вторых, мы предполагаем, что существует способ обработки и нормализации цифрового экземпляра формы, позволяющий сохранить области псевдоизометрии в цифровых экземплярах формы. Например, эти предположения выполняются для формы паспорта гражданина РФ и бытовых сканеров.

*Жесткими документами* будем называть документы, созданные с помощью жестких форм. Все остальные документы являются *гибкими документами*. Для гибкого документа не существует ограничения на расстояния между любой парой соответствующих геометрических объектов в двух экземплярах одного документа.

Известно применение моделей жестких документов для распознавания документов с целью извлечения данных. В них извлекаемые данные описываются однострочными или многострочными полями в локализованном, нормализованном и ректифицированном изображении документа. Границы полей определяются привязкой к опорным элементам. Расстояния от опорных элементов до границ полей определяются в процессе обучения. В [8] опорными элементами являются локальные особенности (ключевые точки), сгруппированные в созвездия точек. Модель состоит из нескольких созвездий. Дескрипторами ключевых точек могут быть SIFT, SURF, YAPE [9]. Для каждого типа дескриптора применяется свой детектор (алгоритм поиска ключевой точки и вычисления дескриптора). Сравнение множества точек модели и найденных точек в анализируемом документе проводится методом RANSAC, который оценивает близость модели и документа с помощью случайных выборок [10]. Формирование случайных выборок становится возможным из-за достаточно большого числа содержащихся в образе документа ключевых точек.

## 2. Различные типы гибких документов

Предлагается различать следующие классы гибких документов:

- гибкие документы, продуцированные одной формой;
- гибкие документы, продуцированные несколькими формами;
- гибкие документы, продуцированные неопределенным числом форм;
- условно-жесткие документы.

Документ с одной формой является жестким. В форме гибкого документа могут меняться:

- шрифты и размеры шрифтов;
- линии подчеркивания;
- разграфка таблиц.

Это приводит к изменению расстояний между *опорными элементами* (словами и линиями разграфки). Распознанные слова являются аналогами точек для жестких форм. В работах [11, 12] описана модель текстовых особых точек. Дескриптором текстовой особой точки является распознанное слово, а детектором – алгоритмы используемой OCR. Для поиска отрезков применятся векторизатор. Отдельный отрезок, равно как и некоторые слова, не являются уникальными (ключевыми) в контексте страницы или фрагмента страницы. Однако в модели [11] для привязки предлагается использовать уникальные в контексте фрагмента страницы комбинированные дескрипторы над словами и отрезками.

Модели гибких документов используют отношения между объектами (словами, отрезками). Отношения между словами, отрезками и полями не могут быть заданы аналогично формуле (1). Но, аналогично (1) для жестких документов, если для пары объектов (A, B) одного изображения справедливо отношение A⊗B, то и для другого изображения этого документа должно выполняться $f(A)\otimes f(B)$.

Совокупность объектов и набор отношений над объектами определяет созвездие. Созвездия являются моделями, для элементов которых выбираются распознанные слова и детектированные графические примитивы. Привязанные созвездия позволяют прогнозировать границы полей. В отличие от отношений между геометрическими ключевыми точками в жестких документах, в гибких документах отношения могут задавать расстояние между *текстовыми особыми точками* (ТОТ, ключевыми словами с координатной рамкой [11]) и окрестностями с учетом существенных искажений. В качестве ⊗ могут выступать следующие отношения:

– $W \in \varepsilon$ – слово W принадлежности окрестности ε, слова в которой линейно упорядочены,

– $W_1 > W_2$ – слово $W_2$ размещено "после" слова $W_1$ для слов из одной окрестности (симметричное отношение – <, "до"),

– $W_1 \vee W_2$ – слово $W_2$ размещено "выше" слова $W_1$ для слов из одной окрестности (симметричное отношение – ∧, "ниже").

Могут применяться следующие метрики:

– $d_1(W_1, W_2)$ – количество слов в промежутке между двумя словами $W_1$ и $W_2$,

– $d_2(W_1, W_2)$ – сумма ширин слов, размещенных между словами $W_1$ и $W_2$,

– $d_x(W_1, W_2)$, $d_y(W_1, W_2)$ – евклидово расстояние между двумя проекциями рамок слова $W_1$ и $W_2$.

Созвездия создаются для того, чтобы описать уникальную в некоторой окрестности последовательность ключевых слов, в которой каждый элемент не обязан быть уникальным для данной окрестности. Рассмотрим созвездие Z, состоящее из упорядоченного набора текстовых особых точек $\{W_1(Z), W_2(Z), \ldots W_n(Z)\}$, удовлетворяющих нескольким условиям:

$W_1(Z) > W_2(Z) > \ldots > W_n(Z)$
$d_1(W_i(Z), W_j(Z)) < \delta^1_{ij}$ или $d_1(W_i(Z), W_j(Z)) > \Delta^1_{ij}$
$d_2(W_i(Z), W_j(Z)) < \delta^2_{ij}$ или $d_2(W_i(Z), W_j(Z)) > \Delta^2_{ij}$
$d_x(W_i(Z), W_j(Z)) < \delta^x_{ij}$ или $d_x(W_i(Z), W_j(Z)) > \Delta^x_{ij}$
$d_y(W_i(Z), W_j(Z)) < \delta^y_{ij}$ или $d_y(W_i(Z), W_j(Z)) > \Delta^y_{ij}$.

Для каждого ключевого слова известен признак, показывающий, является ТОТ обязательной, нейтральной или запрещающей. Например, в описании созвездия

$$D(F_1, F_2, \ldots, F_{p2}) = \begin{cases} W_1^{top};\ W_2^{top};\ \ldots\ W_{p1}^{top}; \\ W_1; \frac{F_1}{S_1};\ W_2; \frac{F_2}{S_2};\ \ldots\ W_{p2}; \frac{F_{p2}}{S_{p2}};\ W_{p2+1}; \\ W_1^{bottom};\ W_2^{bottom};\ \ldots\ W_{p3}^{bottom} \end{cases} \quad (2)$$

используются ключевые слова и фразы

$$W_1^{top};\ W_2^{top};\ \ldots\ W_{p1}^{top};$$
$$W_1;\ W_2;\ \ldots\ W_{p2};\ W_{p2+1};$$
$$W_1^{bottom};\ W_2^{bottom};\ \ldots\ W_{p3}^{bottom}$$

и отрезки $S_1;\ S_2;\ \ldots\ S_{p2}$. Для детектирования отрезков могут быть использовании методы выделения линий и контуров плоских фигур [13, 14, 15]. Для элементов созвездия $D(F_1, F_2, \ldots, F_{p2})$ выполняются следующие условия:

– любая текстовая особая точка $W_1;\ W_2;\ \ldots\ W_{p2}$ размещена ниже, чем любая из точек $W_1^{top};\ W_2^{top};\ \ldots\ W_{p1}^{top}$ и выше, чем любая из точек $W_1^{bottom};\ W_2^{bottom};\ \ldots\ W_{p3}^{bottom}$;

– текстовая особая точка $W_1$ размещена слева от точки $W_2$, точка $W_{p2-1}$ размещена справа от точки $W_{p2}$, оставшиеся точки $W_i$ размещены справа от точки $W_{i-1}$ и слева от точки $W_{i+1}$;

– отрезки $S_i$ размещены справа от точки $W_i$ и слева от точки $W_{i+1}$.

Результатом привязки являются однострочные рамки для полей $F_1;\ F_2;\ \ldots\ F_{p2}$.

Для рассмотренных классов документов детектирование поля заканчивается либо прогнозом одной рамки, либо неудачей.

Привязка созвездий позволяет классифицировать фрагменты распознанного документа. После классификации распознанный деловой документ $D^{REC}$ выглядит следующим образом:

$$D^{REC} = \Theta \cup \eta_1 \cup \eta_2 \cup \ldots,$$

где $\eta_i$ – либо привязанная, либо непривязанная окрестность, Θ – область документа, не подлежащая классификации. С помощью описанного способа классификации мы получаем разбиение распознанного документа на окрестности $\eta_i$. Такое разбиение позволяет упрощать отождествление ТОТ. Например, очевидно, что отождествление слова со словами из строки ηi во всех отношениях существенно проще, чем отождествление слова со всеми распознанными словами документа $D^{REC}$ без анализа структуры.

В общем случае распознается не один документ, а поток документов нескольких типов. В потоке документов могут содержаться как документы известных типов, так произвольные документы. Анализ такого потока упрощается, если классифицировать документы по типам, и далее распознавать документы, для которых известны модели. Однако в реальных потоках число типов документов может составлять $10^3$-$10^4$. Часто необходим анализ документов с большой межклассовой близостью. Близость определяется применяемыми алгоритмами классификации. Одна из причин наблюдаемой близости состоит в незначительных модификациях применяемой формы.

Гибкие документы, продуцированные неопределенным числом форм, иногда называют документами без шаблона. Примерами являются чеки, акты, счета, деловые письма. В реальности для создания документов таких типов также используются формы, но их количество может превышать сотни и тысячи. Например, в документе «счет» указываются следующие реквизиты: сведения о заказчике и продавце, дата, номер, подпись, цена, предмет инвойса. Размещение реквизитов в документе «счет» может существенно варьироваться. Кроме того, часть реквизитов может отсутствовать, а также не иметь опорных элементов.

Общий подход к анализу таких документов заключается в минимизации числа полей, подлежащих распознаванию. Часть проблем решается в системах, обученных на репрезентативных датасетах большого объема. Например, в [16, 17] описана система CloudScan, которая извлекает данные из счета-фактуры рекуррентной нейронной сетью и обеспечивает высокую точность извлечения. В системе CloudScan не используются описания документов счетов-фактур. Так CloudScan обеспечивает извлечение 8 полей. Для обучения использовался датасет из 326471 счетов-фактур [16].

Для документов без шаблона могут быть созданы модели с небольшим числом полей, соответствующих реквизитам. Для этого пригодны описанные выше дескрипторы. Эффективным является поиск полей по их значению. Такой поиск возможен для полей с заполнением с известной структурой текста, например, дат, текста с фиксированным префиксом или суффиксом, штриховых кодов, штампов и т.п.

## 3. Особенности распознавания условно-жестких документов

Одним из классов гибких документов являются условно-жесткие документы, в цифровых экземплярах которых содержится не только одна или несколько областей псевдоизометрии, но и области, в которых не выполняется условие (1). В документы, при создании которых использовалась жесткая форма, в процессе печати и оцифровке в изображение этих документов могут быть внесены неустранимые искажения, не позволяющие выделить области псевдоизометрии. Примерами таких искажений являются сильные замятия страниц, приводящие к сильному искажению формы страницы и областей изображения. Другим примером является применение шрифтов и других статических элементов малого размера, что приводит к значительным потерям точек, которые могли бы быть взяты в качестве ключевых.

Рассмотрим способ привязки полей гибкого документа, описанный в работах [11, 18]. Классификация строк (привязка строк) может проводиться как для всей страницы, так и для отдельной окрестности $\eta$. Заранее задаются модели допустимых строк $M_{s1}, M_{s2}, \ldots, M_{sq}$, каждая модель $M_S$ состоит из набора текстовых особых точек

$$M_S = \{W^+_1, W^+_2, \ldots, W^+_{k+}, W_1, W_2, \ldots, W_k, W^-_1, W^-_2, \ldots, W^-_{k-}\}$$

и параметры, важнейшим из которых является $d_{\mathrm{LINK}}(M_S)$ – пороговое значение числа привязанных точек для надежной привязки. В модели допустимых строк используются три *мешка слов*: запрещенные слова $W^-(S)=\{ W^-_1, W^-_2, \ldots, W^-_{k-}\}$, обязательные слова $W^+(S)=\{ W^+_1, W^+_2, \ldots, W^+_{k+}\}$ и необязательные слова $W(S)=\{ W_1, W_2, \ldots, W_k\}$. Метод привязки распознанных строк $S_w$ состоит в вычислении оценок соответствия моделям $\Delta(S_w, M_{si})$. Оценка $\Delta(S_w, M_{si})$ равняется 0, если была привязана хотя бы одна точка из множества $W^-(S_w)$. Оценка $\Delta(S_w, M_{si})$ также равняется 0, если не было привязано ни одной точки из множества $W^+(S_w)$. Оценка $\Delta(S_w, M_{si})$ равняется 1, если не было привязано ни одной точки $W^-(S_w)$ и было привязано не менее $n_{\mathrm{LINK}} \geq d_{\mathrm{LINK}}(M_{Si})$ точек

$W(S)$ и $W^{+}(S)$. Такая модель в сочетании с предварительной привязкой окрестностей обеспечивает высокую точность привязки строк.

После привязки строк проводится поиск (прогноз) рамок полей для последующего извлечения информации. Для рамки каждого поля F задаются ТОТ, размещенные справа или слева:

- …, $W_{-2}(F)$, $W_{-1}(F)$, F, $W_1(F)$, $W_2(F)$ …– для полей на центральных позициях в строке;
- …, $W_{-2}(F)$, $W_{-1}(F)$, F – для полей на последней позиции в строке;
- F, $W_1(F)$, $W_2(F)$, …– для полей на первой позиции в строке.

Условно-жесткий документ может быть распознан как гибкий документ с одной формой. В общем случае для привязки поля гибкого документа в описанной окрестности необходимо привязка его ближайших *соседей*, то есть текстовых особых точек $W_{-1}(F)$ и $W_1(F)$. Если $W_{-1}(F)$ или $W_1(F)$ не привязаны, но привязаны удаленные соседи $W_{-2}(F)$ или $W_2(F)$, то возможен прогноз рамки F. Точность этого прогноза будет тем меньше, чем больше вариация используемых шрифтов.

Однако для условно-жесткого документа шрифт не меняется и прогноз левой (правой) границы поля F с помощью любой из ТОТ будет … $W_{-2}(F)$, $W_{-1}(F)$ ($W_1(F)$, $W_2(F)$, … Точность прогноза зависит от параметров $\delta$ и $\varepsilon$ в области псевдоизометрии. Разумеется, часть строки, являющаяся окрестностью поля, должна принадлежать области псевдоизометрии.

Привязка последовательности строк в области псевдоизометрии может быть проведена с существенно меньшими требованиями к мешкам $W^{-}(S)$, $W(S)$ и $W^{+}(S)$. В простейшем случае в привязной окрестности $\eta$ в области псевдоизометрии установление соответствия строк $M_{s1}$, $M_{s2}$,…, $M_{sq}$ и $q$ найденных строкам-кандидатов $M^{REC}{}_1$, $M^{REC}{}_2$,… , $M^{REC}{}_q$ является однозначным, вне зависимости от качества распознавания каждой из строк $M^{REC}{}_j$. Уверенность в такой тривиальной привязке основана на следующем:

- надежности привязки окрестности $\eta$;
- малости параметров $\delta$, $\theta$ и $\varepsilon$ области псевдоизометрии.

Обе характеристики могут быть оценены экспериментально на репрезентативном наборе данных.

Аналогично привязка слов в известной строке может проводиться для слов, с большим ожидаемым числом ошибок распознавания. При этом снижение требований к точности распознавания слов возможно при сохранении точности детектирования границ слов. Снижение требований к точности распознавания может принести выигрыш в быстродействии обработки страницы.

Рассмотрим задачу привязки в условно-жестком документе созвездия, состоящего из полей пометок, заданных дескриптором:

$$\mathrm{D}\left(\mathrm{B}_{1,1}, \mathrm{B}_{1,2}, \dots, \mathrm{B}_{1,q1}, \mathrm{B}_{2,1}, \mathrm{B}_{2,2}, \dots, \mathrm{B}_{2,q2,} \dots\right) = \\ = \begin{cases} \mathrm{W}_1^{\mathrm{top}};\ \mathrm{W}_2^{\mathrm{top}}; \dots\ \mathrm{W}_{p1}^{\mathrm{top}}; \\ \mathrm{B}_{1,1}; \mathrm{B}_{1,2};\ \dots\ \mathrm{B}_{1,q1}; \\ \mathrm{B}_{2,1}; \mathrm{B}_{2,2};\ \dots\ \mathrm{B}_{2,q2}; \\ \dots \\ \mathrm{W}_1^{\mathrm{bottom}};\ \mathrm{W}_2^{\mathrm{bottom}}; \dots\ \mathrm{W}_{p3}^{\mathrm{bottom}}. \end{cases} \quad (3)$$

В дескрипторе (3) в отличие от дескриптора (2) не используются ключевые слова для указания границ полей пометок. Текстовые особые точки $\mathrm{W}_1^{\mathrm{top}};\ \mathrm{W}_2^{\mathrm{top}}; \dots\ \mathrm{W}_{p1}^{\mathrm{top}}$ и $\mathrm{W}_1^{\mathrm{bottom}};\ \mathrm{W}_2^{\mathrm{bottom}}; \dots\ \mathrm{W}_{p3}^{\mathrm{bottom}}$ задают окрестность $\eta$, в которой проводится поиск всех или части пометок $B_{i,j}$. Каждая поле-пометка задана координатами $\{\mathrm{B}_{i,j}^{x_1}, \mathrm{B}_{i,j}^{y_1}, \mathrm{B}_{i,j}^{x_2}, \mathrm{B}_{i,j}^{y_2}\}$. Исходными данными для привязки является набор рамок найденных в оцифрованном документе в окрестности $\eta$ *прямоугольников-кандидатов* $\mathrm{B}_\eta = \{\mathrm{b}_1, \mathrm{b}_2, \dots\ \mathrm{b}_q\}$. Каждый прямоугольник-кандидат $b_k$ образован несколькими отрезками, детектированными векторизитором, и описан координатами $\{\mathrm{b}_k^{x_1}, \mathrm{b}_k^{y_1}, \mathrm{b}_k^{x_2}, \mathrm{b}_k^{y_2},\}$. При детектировании возможны несколько классов ошибок:

- ошибки пропуска пометок – к некоторым полям $B_{i,j}$ невозможно привязать ни один прямоугольник-кандидат;\
- ошибки детектирования ложных пометок – не соответствуют никакому полю;
- ошибки размеров – некоторые из прямоугольников-кандидатов $b_k$ соответствуют прообразу поля в оцифрованном документе, однако их размеры исказились из-за зашумления или собственно образа рукописной пометки.

Задача привязки пометок в окрестности условно-жесткого документа состоит в установлении соответствия между множеством полей

и множеством прямоугольников-кандидатов. Если поле $B_{i,j}$ привязано к прямоугольнику $b_t$, то должны соблюдаться соотношения между размерами объектов:

$$\begin{cases} \left| |B_{i,j}^{x_2} - B_{i,j}^{x_1}| - |b_t^{x_2} - b_t^{x_1}| \right| < \varepsilon_x \\ \left| |B_{i,j}^{y_2} - B_{i,j}^{y_1}| - |b_t^{y_2} - b_t^{y_1}| \right| < \varepsilon_y \end{cases} \quad (4)$$

и при этом, если рамка любого другого поля $B_{n,m}$, привязанного к прямоугольнику $b_r$, связана с рамкой одним из заданных выше отношений $B_{i,j} \otimes B_{n,m}$, то это же отношение должно быть выполнено и для рамок привязанных прямоугольников $b_t \otimes b_r$. Перебор прямоугольников-кандидатов при решении задачи привязки упрощается, если задать несколько отношений $\otimes$.

Точность описанного способа привязки полей пометок определяется параметрами области псевдоизометрии, соответствующей найденной окрестности. Параметры отношений $\otimes$, а также $\varepsilon_x$ и $\varepsilon_y$ для условно-жестких документов могут быть определены при обучении на репрезентативном датасете. Для гибких документов это невозможно из-за предположения о произвольности вносимых изменений.

Рассмотрим пример созвездия, состоящего только из пометок. На Рис. 1 обозначены 6 пометок ($B_1$ – $B_6$), сгруппированных в 3 горизонтальные группы. Для этого созвездия мы зададим ориентировочные размеры пометки $\varepsilon_x$ и $\varepsilon_y$, 3 горизонтальные группы пометок ($G_1$, $G_2$, $G_3$) и следующие отношения:

$$\begin{cases} G_1 \vee G_2, G_2 \vee G_3, \\ B_1 \in G_1, B_2 \in G_1, \\ B_3 \in G_2, B_4 \in G_2, \\ B_5 \in G_3, B_6 \in G_3, \\ B_1 < B_2, B_3 < B_4, B_5 < B6, \\ d_x(B_1, B_2) < \delta_x^{12}, d_x(B_1, B_2) > \Delta_x^{12}, d_y(B_1, B_2) < \delta_y^{12}, \\ d_x(B_3, B_4) < \delta_x^{34}, d_x(B_3, B_4) > \Delta_x^{34}, d_y(B_3, B_4) < \delta_y^{34}, \\ d_x(B_5, B_6) < \delta_x^{56}, d_x(B_5, B_6) > \Delta_x^{56}, d_y(B_5, B_6) < \delta_y^{56}. \end{cases} \quad (5)$$

Рассмотрим несколько наборов прямоугольников-кандидатов, найденных для изображений с низким качеством оцифровки (Рис. 2, 3). Фальшивые пометки, обозначенные пунктирными линиями, на Рис. 2 и 3, игнорируются как из-за малого размера (невыполнения условия (4)), так и из-за несоответствия отношений (5) между рамками полей пометок. Отсутствующая пометка $B_5$ на Рис. 3 детектируется с помощью координат привязанной пометки $B_5$ и расстояний $\delta_x^{56}$, $\Delta_x^{56}$ и $\delta_y^{56}$.

Было проведено тестирование предложенного дескриптора для пометок (4) с условиями (5). Использовался собственный тестовый датасет, содержащий 418 изображений условно-жесткого документа. Этот документ включал фрагмент с пометками, изображенными на Рис. 1. Документы были напечатаны на листах размера А4 и оцифрованы камерами мобильных устройств в различных условиях освещения и съемки и были подвергнуты проективным искажениям и нелинейным

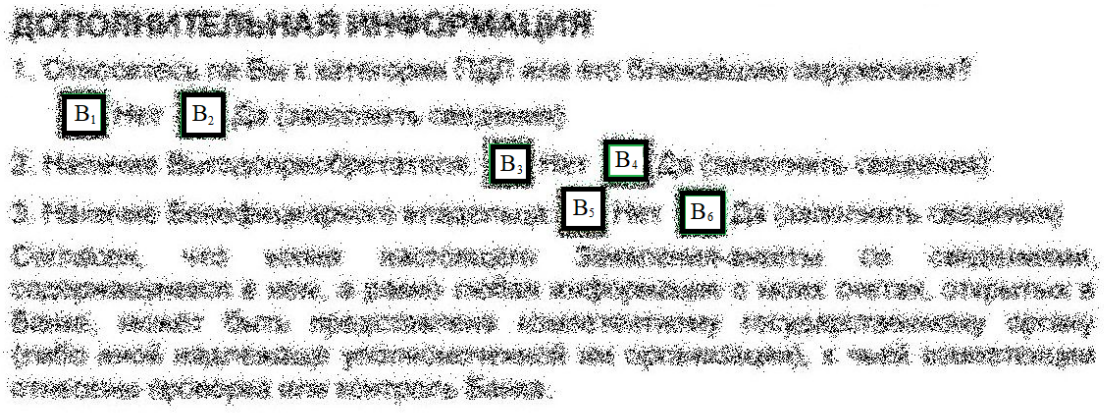


Рис. 1. Пример идеальных пометок в текстовой окрестности

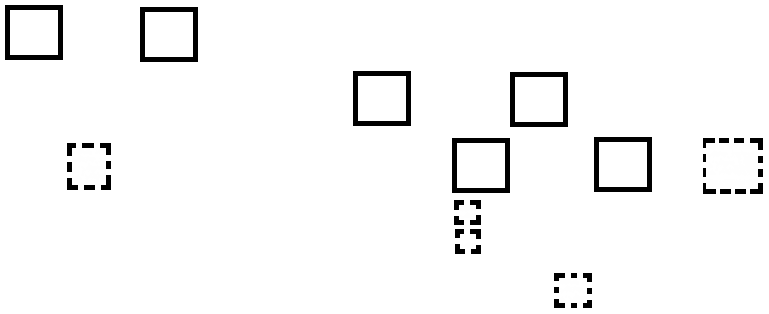

Рис. 2. Примеры найденных прямоугольников-кандидатов с фальшивыми пометками

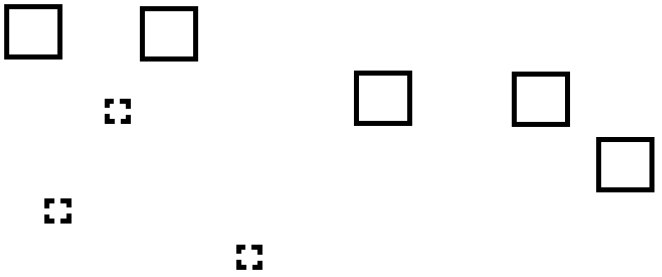

Рис. 3. Примеры найденных прямоугольников-кандидатов с потерянными пометками

деформациям листов. Для распознавания использовался SDK Smart Document Engine [19]. Результаты распознавания приведены в Табл. 1. В результате применения дескриптора (4) средняя точность распознавания одной пометки увеличилась с 87.85% до 88.94%. Другими словами, на тестовом датасете доля ошибок распознавания пометок уменьшилась более чем в 2 раза.

Аналогично можно обеспечить привязку созвездия, состоящего из горизонтальных отрезков, заданных дескриптором:

$$\mathrm{D}\left(\mathrm{F}_{1,1}, \mathrm{F}_{1,2}, \ldots, \mathrm{F}_{1,q2}, \mathrm{F}_{2,1}, \mathrm{F}_{2,2}, \ldots, \mathrm{F}_{2,q2}, \ldots\right) =$$

$$= \begin{cases} \mathrm{W}_1^{\mathrm{top}};\ \mathrm{W}_2^{\mathrm{top}};\ \ldots\ \mathrm{W}_{p1}^{\mathrm{top}}; \\ \frac{\mathrm{F}_{1,1}}{\mathrm{S}_{1,1}}; \frac{\mathrm{F}_{1,2}}{\mathrm{S}_{1,2}};\ \ldots\ \frac{\mathrm{F}_{1,q2}}{\mathrm{S}_{1,p2}}; \\ \frac{\mathrm{F}_{2,1}}{\mathrm{S}_{2,1}}; \frac{\mathrm{F}_{2,2}}{\mathrm{S}_{2,2}};\ \ldots\ \frac{\mathrm{F}_{2,q2}}{\mathrm{S}_{2,q2}}; \\ \ldots \\ \mathrm{W}_1^{\mathrm{bottom}};\ \mathrm{W}_2^{\mathrm{bottom}};\ \ldots\ \mathrm{W}_{p3}^{\mathrm{bottom}}. \end{cases} \quad (6)$$

В дескрипторе (6) поля $\mathrm{F}_{i,j}$ определены с помощью отрезков $\mathrm{S}_{i,j}$, заданных координатами $\{\mathrm{S}_{i,j}^{x_1}, \mathrm{S}_{i,j}^{x_2}, \mathrm{S}_{i,j}^{y}\}$. Исходными данными для привязки является набор рамок найденных в оцифрованном документе в окрестности линий-кандидатов. Набор отношений между линиями созвездия, аналогичный (4), позволяет установить соответствие между множеством полей линий и множеством линий-кандидатов. Точность привязки полей созвездия также определяется параметрами области псевдоизометрии, соответствующей найденной окрестности.

## 4. Создание дескрипторов для условно-жестких документов

Описанный в данной работе подход основан на созвездиях текстовых особых точек и других объектов (пометки и отрезки). Самое простое созвездие состоит из одной единственной текстовой особой точки. Другим примером является *шингл* – последовательность слов, размещенных одно за другим:

$$\mathrm{W}_1(\mathrm{Z}) > \mathrm{W}_2(\mathrm{Z}) > \ldots > \mathrm{W}_n(\mathrm{Z})$$
$$d_1(\mathrm{W}_{i+1}(\mathrm{Z}), \mathrm{W}_i(\mathrm{Z})) < 1.$$

Многие фрагменты документа могут быть описаны с помощью *цепей*, состоящих из последовательности слов $\mathrm{W}_1(\mathrm{Z}) > \mathrm{W}_2(\mathrm{Z}) > \ldots > \mathrm{W}_n(\mathrm{Z})$. Пары соседних слов $(\mathrm{W}_i(\mathrm{Z}), \mathrm{W}_{i+1}(\mathrm{Z}))$ образует битерм, для которого может быть задано ограничение на расстояние между словами битерма:

$$d_1(\mathrm{W}_i(\mathrm{Z}), \mathrm{W}_{i+1}(\mathrm{Z})) < \delta^2{}_i.$$

Табл. 1. Результаты распознавания условно жесткого документа

| Пометка | Точность привязки (%) | | Точность распознавания (%) | |
|---|---|---|---|---|
| | без дескриптора (4) | с дескриптором (4) | без дескриптора (4) | с дескриптором (4) |
| $B_1$ | 98.79 | 98.96 | 85 | 88.33 |
| $B_2$ | 98.22 | 98.39 | 89.03 | 90.02 |
| $B_3$ | 99.26 | 99.31 | 90.52 | 91.27 |
| $B_4$ | 98.57 | 98.66 | 83.29 | 84.04 |
| $B_5$ | 99.35 | 99.39 | 87.28 | 88.28 |
| $B_6$ | 99.00 | 99.05 | 91.27 | 91.52 |

Найденный фрагмент задает окрестность, область которой существенно меньше площади страницы документа. Привязку полей в найденной окрестности, являющейся областью псевдоизометрии, можно проводить в предположении существенно большего числа заданных отношений между полями и опорными элементами.

Процесс обучения состоит из нескольких этапов. Первый этап состоит в выборе шинглов, на основе которых можно классифицировать образы документов и разбивать образы документов на окрестности. Для рассмотренных нами более 100 типов гибких и условно-жестких деловых документов для классификации фрагментов было достаточно шинглов и цепей.

Второй этап – это построение модели, то есть выбор слов, образующих цепи, которые имеются в одном документе (фрагменте) и отсутствуют в других документах (фрагментах). Цепи классифицируют распознанные тесты существенно более надежно, чем отдельные слова. На втором этапе имеет смысл построить модели на оцифрованных формах без заполнения. Обучение может быть проведено методами ML или вручную.

Третий этап состоит в настройке параметров созвездий, то есть в установлении порогов между некоторыми ключевыми словами для некоторых метрик. Третий этап следует проводить на образцах с заполнением. Основной трудностью при обучении является создание датасетов. Общедоступных тестовых наборов данных не существует не только для идентификационных документов [20], но и для печатных форм организаций.

Для условно-жестких документов для опорных элементов необходимо задать набор отношений, их связывающих. При настройке параметров созвездий условно-жестких документов необходимо оценить области псевдоизометрии и их параметры. Возможен случай, когда для некоторого поля невозможно на всех экземплярах датасета найти область псевдоизометрии с разумными параметрами. Это означает, что для данного поля следует применять только дескрипторы, основанные на отношениях и метриках для гибких документов.

И, наоборот, при обучении гибких документов имеет смысл детектировать области псевдоизометрии и использовать отношения между объектами аналогично (4).

## Заключение

В работе приведена классификация гибкости документов: жесткие документы, условно-жесткие документы, гибкие документы, продуцированные одной формой, гибкие документы, продуцированные малым и большим числом форм. Наибольшая точность извлечения данных достигается для жестких документов. Точность извлечения данных из гибких документов зависит не от сложности модели, а от структуры документа, качества оцифровки, возможностей улучшения изображений.

В работе предложено понятие области псевдоизометрии. Дескрипторы, построенные в предположении успешного детектирования области псевдоизометрии позволяют для условно-жестких документов применять такие отношения между опорными элементами и полями, что привязка полей будет возможна даже в случае очень низкой точности распознавания слов статического текста. Области псевдоизометрии имеет смысл детектировать и при обучении гибких документов.

Описанные эксперименты показывают, что для некоторых полей пометок в условиях сильного зашумления и значительных искажений доля ошибок уменьшается в два раза.

## Литература


1. Rusiñol M., Frinken V., Karatzas, D., Bagdanov, A. D., Lladós, J.: Multimodal page classification inadministrative document image streams. In: IJDAR. 17(4), 331–341 (2014). https://doi.org/10.1007/s10032-014-0225-8
2. Postnikov V. V.: Identification and Recognition of Documents with a Predefined Structure // Pattern Recognition and Image Analysis. 13(2), 332–334 (2003)
3. Jain, R., Wigington, C.: Multimodal Document Image Classification. 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. 142–147 (2019). https://doi.org/10.1109/ICDAR.2019.00031
5. Vasiliev, S.S., Korobkin, D.M., Kravets, A.G., Fomenkov, S.A., Kolesnikov, S.G.: Extraction of cyber-physical systems inventions' structural elements of russian-language patents. Stud. Syst. Springer, Decis. Control, 259, 55–68 (2020). https://doi.org/10.1007/978-3-030-32579-4_5
6. 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
7. Augereau, O., Journet, N., Domenger, J.-P.: Semi-structured document image matching and recognition/ IS&T/SPIE Electronic Imaging., 13–24 (2013). https://doi.org/10.1117/12.2003911
8. Skoryukina, N, Arlazarov, V, Nikolaev, D.: Fast Method of ID Documents Location and Type Identification for Mobile and Server Application. IEEE International Conference on Document Analysis and Recognition (ICDAR): 850–857 (2019). https://doi.org/10.1109/ICDAR.2019.00141
9. Bellavia, F.: SIFT Matching by Context Exposed. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022). https://doi.org/10.1109/TPAMI.2022.3161853
10. Skoryukina, N., Faradjev, I., Bulatov, K;, Arlazarov, V.: Impact of geometrical restrictions in RANSAC sampling on the ID document classification. Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2020), 1143306R (2020). https://doi.org/10.1117/12.2559306
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. 350, 173–184. (2021). https://doi.org/10.1007/978-3-030-67892-0_15
12. Bay, H., Tuytelaars, T., Van Gool, Luc.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding - CVIU. 110(3), 404–417 (2006).
13. Matas, J., Galambos, C., Kittler, J.: Robust Detection of Lines Using the Progressive Probabilistic Hough Transform, Computer Vision and Image Understanding, 78(1), 119–137 (2000). https://doi.org/10.1006/cviu.1999.0831
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. 32(4), 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, 417, 243–246 (2022). https://doi.org/10.1007/978-3-030-95116-0_20
16. Palm, R. B., Winther, O., Laws F.: CloudScan - A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 406-413 (2017). https://doi.org/10.1109/ICDAR.2017.74
17. Pegu, B., Singh, M., Agarwal, A., Mitra, A., Singh, K.: Table Structure Recognition Using CoDec Encoder-Decoder. In: Barney Smith, E.H., Pal, U. (eds) Document Analysis and Recognition – ICDAR 2021 Workshops. Lecture Notes in Computer Science, 12917, 66-80 (2021). https://doi.org/10.1007/978-3-030-86159-9_5
18. Slavin, O. A.: Using Special Text Points in the Recognition of Documents. Studies in Systems, Decision and Control. Springer Nature Switzerland AG., 259, 43–53 (2020). https://doi.org/10.1007/978-3-030-32579-4_4
19. Smart Document Engine – automatic analysis and data extraction from business documents for desktop, server and mobile platforms. https://smartengines.com/ocr-engines/document-scanner. Last access 16 may 2023
20. Awal, A.M., Ghanmi, N., Sicre, R., Furon, T.: Complex Document Classification and Localization Application on Identity Document Images. Proc. 14th IAPR International Conference on Document Analysis and Recognition. 427-432 (2017). https://doi 10.1109/ICDAR.2017.77


**Славин Олег Анатольевич**. Федеральное государственное учреждение Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, Россия. Главный научный сотрудник, доктор технических наук, доцент. Область научных интересов: распознавание образов, информационные системы. E-mail: oslavin@isa.ru.

# Field Binding Algorithms for Recognizing Conditionally Rigid Administrative Documents

O. A. Slavin

Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia
LLC "Smart Engine Service", г. Moscow, Russia

**Abstract**. Definitions of flexible and rigid documents used in technologies for entering administrative documents into a computer are proposed. The features of creating, digitizing and analyzing rigid forms and rigid documents are considered. The limits of applicability of the model for linking images of rigid documents distorted during digitization are described. A model for linking flexible documents is considered, based on recognized words and graphic primitives connected by a set of order relations. The classification is based on various methods of preparing administrative documents for printing. The features of field binding and recognition for several types of documents are described, such as conditionally rigid documents, flexible documents produced by one form, flexible documents produced by a small and large number of forms. The case of recognition of conditionally rigid documents using flexible document input technologies is considered. The experiments carried out show that for some fields of notes in conditions of strong noise and significant distortion, the proportion of errors is reduced by half.
**Keywords**: document recognition, conditionally rigid document, text feature point, check point.

**DOI** 10.14357/20718632230404 **EDN** BXZTGE

## References

1. Rusiñol M., Frinken V., Karatzas, D., Bagdanov, A. D., Lladós, J.: Multimodal page classification inadministrative document image streams. In: IJDAR. 17(4), 331–341 (2014). https://doi.org/10.1007/s10032-014-0225-8
2. Postnikov V. V.: Identification and Recognition of Documents with a Predefined Structure // Pattern Recognition and Image Analysis. 13(2), 332–334 (2003)
3. Jain, R., Wigington, C.: Multimodal Document Image Classification. 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. 142–147 (2019). https://doi.org/10.1109/ICDAR.2019.00031
5. Vasiliev, S.S., Korobkin, D.M., Kravets, A.G., Fomenkov, S.A., Kolesnikov, S.G.: Extraction of cyber-physical systems inventions’ structural elements of russian-language patents. Stud. Syst. Springer, Decis. Control, 259, 55–68 (2020). https://doi.org/10.1007/978-3-030-32579-4_5
6. 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
7. Augereau, O., Journet, N., Domenger, J.-P.: Semi-structured document image matching and recognition/ IS&T/SPIE Electronic Imaging., 13–24 (2013). https://doi.org/10.1117/12.2003911
8. Skoryukina, N, Arlazarov, V, Nikolaev, D.: Fast Method of ID Documents Location and Type Identification for Mobile and Server Application. IEEE International Conference on Document Analysis and Recognition (ICDAR): 850–857 (2019). https://doi.org/10.1109/ICDAR.2019.00141
9. Bellavia, F.: SIFT Matching by Context Exposed. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2022). https://doi.org/10.1109/TPAMI.2022.3161853
10. Skoryukina, N., Faradjev, I., Bulatov, K;, Arlazarov, V.: Impact of geometrical restrictions in RANSAC sampling on the ID document classification. Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2020), 1143306R (2020). https://doi.org/10.1117/12.2559306
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. 350, 173–184. (2021). https://doi.org/10.1007/978-3-030-67892-0_15
12. Bay, H., Tuytelaars, T., Van Gool, Luc.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding - CVIU. 110(3), 404–417 (2006).
13. Matas, J., Galambos, C., Kittler, J.: Robust Detection of Lines Using the Progressive Probabilistic Hough Transform, Computer Vision and Image Understanding, 78(1), 119–137 (2000). https://doi.org/10.1006/cviu.1999.0831
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. 32(4), 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, 417, 243–246 (2022). https://doi.org/10.1007/978-3-030-95116-0_20
16. Palm, R. B., Winther, O., Laws F.: CloudScan - A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 406-413 (2017). https://doi.org/10.1109/ICDAR.2017.74
17. Pegu, B., Singh, M., Agarwal, A., Mitra, A., Singh, K.: Table Structure Recognition Using CoDec Encoder-Decoder. In: Barney Smith, E.H., Pal, U. (eds) Document Analysis and Recognition – ICDAR 2021 Workshops. Lecture Notes in Computer Science, 12917, 66-80 (2021). https://doi.org/10.1007/978-3-030-86159-9_5
18. Slavin, O. A.: Using Special Text Points in the Recognition of Documents. Studies in Systems, Decision and Control. Springer Nature Switzerland AG., 259, 43–53 (2020). https://doi.org/10.1007/978-3-030-32579-4_4
19. Smart Document Engine – automatic analysis and data extraction from business documents for desktop, server and mobile platforms. https://smartengines.com/ocr-engines/document-scanner. Last access 16 may 2023
20. Awal, A.M., Ghanmi, N., Sicre, R., Furon, T.: Complex Document Classification and Localization Application on Identity Document Images. Proc. 14th IAPR International Conference on Document Analysis and Recognition. 427-432 (2017). https://doi 10.1109/ICDAR.2017.77

**Slavin Oleg A.** Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia. Lead researcher, Doctor of Technical Sciences, Docent. Scientific interests: artificial intelligence, machine learning, recognition systems, information technology. E-mail: oslavin@isa.ru