Поиск несоответствий между отсканированными копиями деловых документов

О. А. Славин
ФГУ ФИЦ «Информатика и управление» РАН
oslavin@isa.ru

Е. И. Андреева[1,2]
[1]ООО «Смарт Энджинс Сервис»
[2]ФГАОУ ВО «Московский физико-технический институт (государственный университет)»
andreeva@phystech.edu

***Аннотация*. В работе рассмотрены методы решения задачи сравнения оцифрованных копий деловых документов. Такая задача возникает при сравнении двух экземпляров документов, подписанных двумя сторонами с целью найти возможные модификации, внесенные одной стороной, например, в банковской сфере при заключении договоров в бумажной форме. Предложен способ сравнения двух оцифрованных изображений на основе комбинирования методов сравнения при применении алгоритмов распознавания текста и методов сравнения сегментированных частей изображения с использованием особых точек. Предложенный метод классификации может быть применен в современных САПР для анализа содержимого текстовых документов.**

Ключевые слова: сравнение документов; сегментация изображений; особые точки; автоматическое распознавание текста; расстояние Левенштейна

I. Введение

При заключении договоров в бумажной форме может возникнуть ситуация, когда одной из сторон могут быть внесены изменения в свой экземпляр договора, которые нежелательны для противоположной стороны, например, в случае изменения клиентом банка условий договора в свою пользу. Чтобы этого избежать, необходимо процедура сравнения документов, цель сравнения состоит в поиске модификаций в тексте документа.

В качестве объектов для сравнения образов деловых документов рассматривались отсканированные страницы, состоящие из строк, слов и отдельных символов.

Мы рассматривали следующие виды возможных модификаций:

- замену одного или нескольких символов в слове;
- замену одного слова на другое;
- добавление слова или группы слов;
- удаление слова или группы слов.

В работе описан метод, который позволяет детектировать перечисленные модификации в отсканированных изображениях деловых документов.

II. Метод

Рассмотрим задачу поиска несоответствий между отсканированными изображениями двух экземпляров страниц делового документа. Одно изображение соответствует эталонной странице, другое – тестовой. Входными данными является эталонное изображение с разметкой на слова. В тестовом изображении необходимо выделить внесенные модификации.

Определим документ как набор строк $D = \left\{ L_i^d \right\}_{i=1}^{|D|}$, строку как набор слов $L^d = \left\{ W_j^l \right\}_{j=1}^{|L^d|}$, а слово как прямоугольник на растре $W^l =< x, y, w, h >$, где (x, y) – координаты левой верхней точки этого прямоугольника, w – его ширина, а h – его высота, W^l – слово на строке l, L^d – строка в документе d.

Два документа считаются свободными от модификаций, если все строки в них скоординированы, скоординированность строк проводится с помощью проверки скоординированности слов, составляющих строки. Определение скоординированных слов дано ниже.

Метод поиска несоответствий состоит из следующих этапов:

- обработка тестового изображения:
  - предварительная обработка изображения,
  - сегментация текста на строки,
  - сегментация текста на слова,
- сравнение строк эталонного и тестового изображений, установление соответствий между словами и строками.

Две строки считаются скоординированными, если доля скоординированных слов в этих строках больше заданного порога $line\_simil$.

Важным этапом предварительной обработки отсканированных изображений являлась нормализация

изображения (доворот), после которой строки текста оказывались в строго горизонтальном положении. Для этого использовался алгоритм быстрого преобразования Хафа [3].

Далее проводится сегментация построением гистограммы, находятся рамки слов на изображении. Сначала проводится выделение компонент связности, выбираются основные компоненты, формируются горизонтальные и вертикальные гистограммы с учетом возможных выбросов. Вертикальная сегментация на строки и горизонтальная сегментация полученных строк на слова проводится с помощью анализа гистограмм и морфологических операций [10]. После сегментации изображения были представлены в виде набора координат рамок слов.

Для сравнения слов используется комбинация нескольких методов.

Первый метод основывается на распознавании текста, полученного в результате применения соответствующих алгоритмов к отсканированным изображениям. Нами использовались два средства распознавания:

- свободно распространяемая программа распознавания текстов OCR Tesseract;
- библиотека распознавания, использованная в программных продуктах, например, в Smart ID Reader [6].

В работе [2] перечислены достоинства OCR Tesseract: возможность свободного распространения и представление результатов распознавания в формате HOCR (HTML OCR), который содержит информацию о координатах рамок распознанных слов. Характерные для OCR Tesseract (мы использовали версию 4.0) ошибки, такие как неверное распознавание структуры страницы и ошибки распознавания символов, затрудняют сравнение документов и вносят ложные срабатывания в результаты сравнения.

Библиотека распознавания Smart Engines [7], основные принципы работы которой описаны в работе [8], обладая меньшей функциональностью по сравнению с OCR Tesseract, обеспечивает более точное распознавание документов, в первую очередь, русскоязычных. Распознавание в этой библиотеке основано на двух типах нейронных сетей. Первая нейронная сеть состоит из нескольких сверточных слоев и двух полносвязных слоев. За каждым сверточным слоем следует слой субдискретизации, а за каждым слоем субдискретизации и полносвязным слоем – эвристический слой случайного частичного обнуления. В качестве функций активации использовался гиперболический тангенс. С целью увеличения точности сегментации и распознавания символов было применено комбинирование сверточной сети и двусторонней рекуррентной нейронной сети. Были использованы рекуррентные нейронные сети архитектуры долгой краткосрочной памяти [9], эффективные для анализа последовательностей различных объектов, в частности для распознавания печатного и рукописного текстов.

Рассмотрим сравнение слов, представленных в виде текстовых строк. С помощью расстояния Левенштейна [1] $lev_{A,B}(A,B)$ определим для двух слов коэффициент сходства следующим образом

$$coeff_{OCR}(A,B) = 1 - lev_{A,B}(A,B) / max(|A|,|B|),$$

где $max(|A|,|B|)$ – максимум из длин слов A и B.

Два слова будем считать скоординированными, если $coeff_{OCR}(A,B)$ больше заданного порога $word\_ocr\_simil$.

Для сравнения распознанных слов необходимо применить следующие операции, связанные с особенностями распознавания OCR Tesseract:

- игнорирование регистра, так как одной из частых ошибок распознавания является изменение регистра случайных букв внутри слова;
- отождествление некоторых символов, таких как короткое и длинное тире, дефис и символ знака минуса, различные виды кавычек, так как при распознавании они могут заменяться друг на друга;
- игнорирование пунктуации, так как возможные ошибки несущественны при сравнении двух документов;
- отождествление символов, являющихся сходными с точки зрения механизма распознавания, например, такие как буква О и цифра 0, буква З и цифра 3 [11].

При работе описанного метода происходит большое число ложных срабатываний, которое можно уменьшить с помощью другого метода сравнения слов, основанного на сопоставлении особых точек.

Для эталонного и тестового изображений детектировались особые точки и для найденных особых точек вычислялись RFD дескрипторы [5]. Далее для нахождения оптимального преобразования используется алгоритм, основанный на методе RANSAC [4].

Вычислялись дескрипторы точек, расположенных в эталонном изображении, на преобразованном тестовом изображении, после чего сравнивались два набора бинарных дескрипторов точек с геометрически близким расположением. После сравнения проводилась очистка от выбросов и рассчитывался коэффициент подобия:

$$coeff_{fp} = 1 - |outliers| / (|outliers| + |inliers|),$$

где $|outliers|$ – количество выбросов, $|inliers|$ – количество не-выбросов. Два слова считались скоординированными, если $coeff_{fp}$ для них больше заранее заданного порога $word\_fp\_simil$.

Результаты комбинирования методов, основанных на распознавании и на особых точках, состоят в найденных соответствиях слов эталонного и тестового изображений. Число ложных срабатываний в результатах сравнения может быть уменьшено анализом слов, являющихся соседями нескоординированных слов.

Производилась проверка склеивания двух слов и разделения слова на несколько частей, которые являются результатом неправильной сегментации. Примером является случай, когда большая часть разделенного слова будет поставлена в соответствие эталонному слову, а меньшая часть будет считаться вставленным словом.

На рисунках ниже приведены несколько случаев ложных срабатываний. На рис. 1–6 красным помечены слова, которым не найдено соответствие, и синим, если тестовое слово скоординировано со словом из эталонной страницы, верхнее слово соответствует эталонному документу, а нижнее – тестовому.

Рис. 1. Тестовое слово разбито на части, соответствие не установлено ни для эталонного слова, ни для частей тестового слова

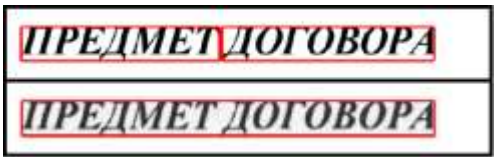

Рис. 2. Два тестовых слова склеились, соответствие не установлено ни для эталонного слова, ни для частей тестового слова

Рис. 3. Два тестовых слова склеились, соответствие не установлено для второй части эталонного слова, остальные слова скоординированы

Рис. 4. Тестовое слово разбито на части, при этом скоординировано эталонное слово и первое из частей тестового слова, а остальным частям тестового слова соответствие не найдено

Рис. 5. Два тестовых слова склеились, соответствие не установлено для первой части эталонного слова, остальные слова скоординированы

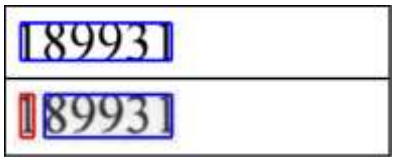

Рис. 6. Тестовое слово разбито на части, при этом скоординировано эталонное слово и второе из частей тестового слова, а первой части тестового слова соответствие не найдено

Для таких случаев уточнение координации слов проводилось путем объединения или разбиения соответствующих частей слов.

Также производилась проверка на смещение слов по строкам (пример приведен на рис. 6). При модификациях типа вставки и удаления слов, слова могут смещаться или переноситься на другие строки, что мешает построчному сравнению, так как перенесенные слова будут считаться модификацией, как показано на рисунке. Чтобы избежать таких ложных срабатываний, происходит проверка на смещение слов по строкам. Для этого при обнаружении смещения рассматривается следующая и предыдущая строка и происходит сопоставление слов на них, независимо от положения на строке.

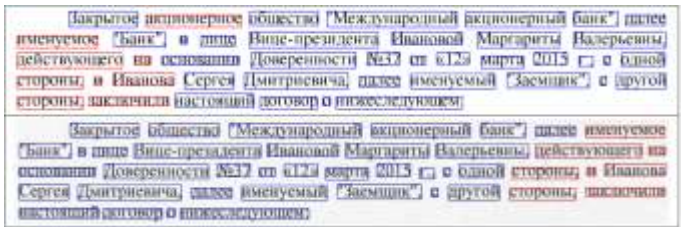

Рис. 7. Пример смещения слов. Верхняя часть рисунка соответствует эталонному документу, а нижняя – тестовому. Видно, что присутствует модификация удаление слова “акционерное”, из-за чего последующие слова смещаются по строкам. Синим обозначены слова, которым найдено соответствие, а красным – те, которым не найдено

III. Результаты

Для экспериментов был подготовлен собственный тестовый набор данных, состоящий из 210 пар документов, состоящих из примерно семидесяти тысяч слов, с 820 модификациями, в том числе вставки и удаления строк. Половина тестовых изображений была отсканирована с разрешением 200 dpi, а другая половина – 300 dpi. Большинство изображений не содержали таблицы и графические элементы.

Тестирование алгоритма состояло из подсчета:

- числа правильно найденных модификаций tp ;
- числа ложных срабатываний fp ;
- число реальных ненайденных модификаций fn ,

и определения характеристик точности и полноты:

$$Precision = tp / (tp + fp)$$

$$Recall = tp / (tp + fn)$$

В таблице I представлены результаты, полученные при использовании только OCR Tesseract, в табл. II – результаты эксперимента при комбинировании результатов распознавания OCR Tesseract с OCR Smart Engines.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ (OCR TESSERACT)

	Precision	**Recall**
300 dpi	64,62%	96,36%
200 dpi	64,14%	98,00%

ТАБЛИЦА II РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ (OCR TESSERACT И OCR SMART ENGINES)

	Precision	**Recall**
300 dpi	90,91%	98,18%
200 dpi	74,38%	98,00%

Как видно из таблиц I и II, использование OCR Smart Engines значительно повышает точность определения модификаций: на 10% для изображений 200 dpi и на 26% для изображений 300 dpi, при этом на полноту обнаружения модификаций практически не влияет. Итоговая полнота обнаружения модификаций достаточно велика и составляет 98% для изображений 300 dpi и 200 dpi. Точность снижается из-за ложных срабатываний, возникающих из-за ошибок распознавания. Ухудшение точности для случая 200 dpi также связано с увеличением количества ошибок распознавания, примеры которых приведены на рис. 8.

Рис. 8. Примеры характерных ошибок. Слева представлены изображения слов, справа результаты распознавания

IV. ВЫВОДЫ

Предложенная в работе комбинация нескольких методов позволяет решить задачу поиска несоответствий между отсканированными копиями деловых документов. Проведенные эксперименты показывают, что на изображениях, отсканированных с разрешением 300 dpi, была достигнута точность 90,9%, полнота составила 98,2%, для изображений, отсканированных с разрешением 200 dpi, точность составила 74,4%, а полнота – 98,0%.

СПИСОК ЛИТЕРАТУРЫ

[1] Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов. М.: Доклады АН СССР, Т. 163, №4, 1965. 845-848 с.

[2] Славин О.А. Метод классификации распознанных страниц деловых документов на основе метода template matching // Труды Седьмой Международной конференции "Системный анализ и информационные технологии" (САИТ – 2017). 2017. С. 667–671.

[3] Nikolaev D., Karpenko S., Nikolaev I., and Nikolayev P. Hough transform: underestimated tool in the computer vision field // Proc. of the 22th European Conference on Modelling and Simulation. 2008. P. 238–246.

[4] Fischler M.A., Bolles R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm. Of the ACM 24: 381–395. June 1981. DOI:10.1145/358669.358692.

[5] Шемякина Ю.А., Жуковский А.Е., Фараджев И.А. Исследование алгоритмов вычисления проективного преобразования в задаче наведения на планарный объект по особым точкам. М.: Искусственный интеллект и принятие решений, № 1. 2017. С. 43-49.

[6] http://smartengines.biz/smart-id-reader/

[7] Библиотека для распознавания идентификационных карт личности "Smart IDReader": свидетельство о государственной регистрации программы для ЭВМ № 2016616961 / Арлазаров В.В., Николаев Д.П., Усилин С.А., Булатов К.Б., Чернов Т.С., Слугин Д.Г., Ильин Д.А., Безматерных П.В., Муковозов А.А., Лимонова Е.Е., № 2016612014; заявл. 10.03.2016; зарегистрировано в реестре программ для ЭВМ 22.06.2016. [1] с.

[8] Чернов Т.С., Ильин Д.А., Безматерных П.В., Фараджев И.А., Карпенко С.М. Исследование методов сегментации изображений текстовых блоков документов с помощью алгоритмов структурного анализа и машинного обучения. Вестник РФФИ, № 4 (92) октябрь - декабрь 2016. С. 55-71. DOI: 10.22204/2410-4639-2016-092-04-55-71

[9] Tang Z., Wang D., Zhang Z. Recurrent neural network training with dark knowledge transfer. // IEEE international conference on acoustics, speech and signal processing (ICASSP) – 2016. P. 5900–5904

[10] Слугин Д.Г., Арлазаров В.В. Поиск текстовых полей документа с помощью методов обработки изображений // М: Труды ИСА РАН, Т. 67 № 4. 2017. С. 65-73

[11] 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. 2015. Т. 65. № 3. С. 85-94.