

Cognitive придумал технологию, которая облегчит подделку документов

14.04.11, Чт, 17:09, Мск, Текст: Редакция

<http://corp.cnews.ru/news/top/index.shtml?2011/04/14/436431>

Cognitive Technologies намерена вывести на рынок свою технологию оцифровки документов ScanPack, которую отличает от аналогов способность распознавать текст с испорченным фоном. Система обладает интересным побочным эффектом: с ее помощью легко выделить в отдельный слой печати и подписи и использовать их для фальсификации документов.

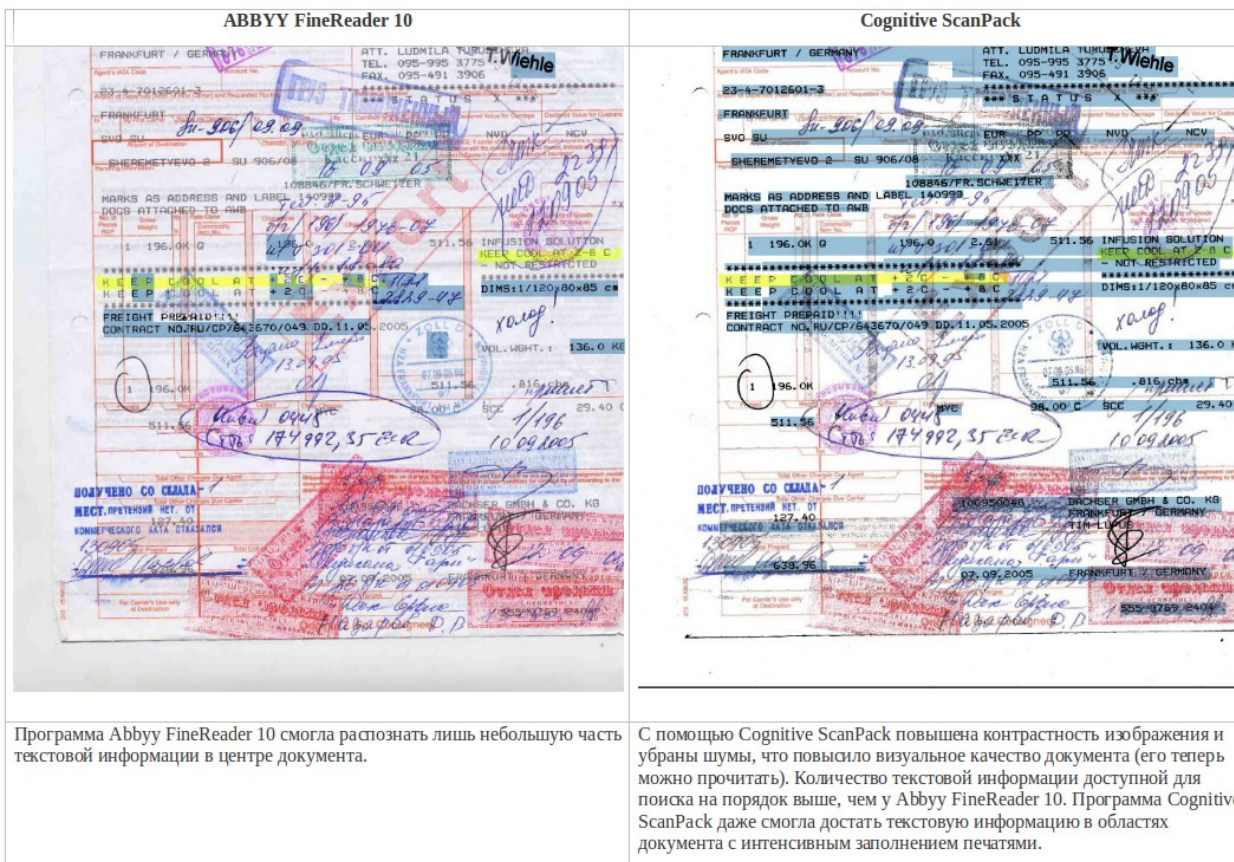
Отечественный разработчик Cognitive Technologies объявил о скором выводе на рынок системы обработки деловых документов с редкими свойствами - Cognitive ScanPack. Система предназначена для решения традиционных задач: сканирования, обработки и сжатия документов. Однако в ней применены несколько технологических особенностей, которые принципиально отличают ее от существующих аналогов.

Принципиальной новацией ScanPack стали новые алгоритмы анализа изображений. Самыми значимыми свойствами, которые благодаря им появились в системе, в компании называют работу с документами со сложной структурой (большим числом печатей поверх текста) и с испорченным фоном («запачканными, ветхими, с залитой машинным маслом конструкторской документацией»). После обработки в ScanPack документ возвращается в «приемлемом визуальном качестве».

Использование для архивации формата PDF/A позволяет сжать исходный документ до 4-10 раз. В итоговом файле возможен текстовый поиск.

ScanPack автоматизирует процесс оцифровки документации от этапа сканирования до сжатия. В Cognitive полагают, что свойства системы делают ее особенно подходящей для работы с деловыми документами.



Григорий Липич, гендиректор Abbyy Россия, оказался не готов дать оценку технологии Cognitive ScanPack до тестирования, однако, заявил, что «подобные технологии существуют на рынке уже давно». Он говорит, что в продуктах его компании используется технология MRC (Mixed Raster Content), которая позволяет значительно уменьшать размер PDF-файлов и получать итоговые документы небольшого размера с возможностью полнотекстового поиска и сохранением первоначального внешнего вида. Она реализована в инструментари для разработчика Abbyy FineReader Engine и в системах потокового ввода документов и данных Abbyy FlexiCapture и Abbyy Recognition Server.



Сравнительный тест ABBYY FineReader и Cognitive ScanPack. Распознанный текст удалось выделить фиолетовым цветом. (Тест проведен и предоставлен кафедрой инженерной кибернетики НИТУ МИСиС)

При использовании технологии MRC изображение перед сжатием проходит через операцию, называемую «разделение на слои»: в изображении выделяются структурные элементы трех типов (текст, изображения (фото, схемы, диаграммы и пр.) и области, залитые одним и тем же цветом). В дальнейшем эти «слои» обрабатываются алгоритмами сжатия, независимо друг от друга.

Кроме того, в решениях Abbyy применяется технология адаптивного распознавания ADRT (Adaptive Document Recognition Technology), которая позволяет обрабатывать документы со сложным форматированием.

| ABBYY FineReader 10 | Cognitive ScanPack |
|---|---|
| <p style="text-align: center;">С М Е Т А № 4 от 31.01 2007г.</p> <p style="text-align: center;">на производство ремонтных работ</p> <p>Настоящая смета составлена ООО «Банарт» в лице Ген. Директора Якимова Александра Борисовича, в дальнейшем именуемого «Подрядчик» и</p> <p>в дальнейшем именуемого «Заказчик», в том, что «Подрядчик» обязуется выполнить ремонтные работы на объекту по адресу _____</p> <p>В полном объеме и по срокам указанном в договоре № _____ от _____ 2007г. По ценам указанным в настоящей смете. Смета составлена в 2-х экземплярах с приложениями на 8 листах. Всекие изменения в смете после ее подписания недопустимы. Изменения происходящие в ходе работ вносятся в приложение к смете (лист «Приложения») и подписываются сторонами. Данная смета согласована сторонами _____ 2007г. Общая сметная стоимость проводимых работ составляет <u>202 733 руб. (двести две тысячи семьдесят три рубля)</u></p> <p>«ЗАКАЗЧИК» _____ «ПОДРЯДЧИК» Ген. директор ООО «Банарт» Якимов</p>  | <p style="text-align: center;">С М Е Т А № 4 от 31.01 2007г.</p> <p style="text-align: center;">на производство ремонтных работ</p> <p>Настоящая смета составлена ООО «Банарт» в лице Ген. Директора Якимова Александра Борисовича, в дальнейшем именуемого «Подрядчик» и</p> <p>в дальнейшем именуемого «Заказчик», в том, что «Подрядчик» обязуется выполнить ремонтные работы на объекту по адресу _____</p> <p>В полном объеме и по срокам указанном в договоре № _____ от _____ 2007г. По ценам указанным в настоящей смете. Смета составлена в 2-х экземплярах с приложениями на 8 листах. Всекие изменения в смете после ее подписания недопустимы. Изменения происходящие в ходе работ вносятся в приложение к смете (лист «Приложения») и подписываются сторонами. Данная смета согласована сторонами _____ 2007г. Общая сметная стоимость проводимых работ составляет <u>202 733 руб. (двести две тысячи семьдесят три рубля)</u></p> <p>«ЗАКАЗЧИК» _____ «ПОДРЯДЧИК» Ген. директор ООО «Банарт» Якимов</p>  |
| <p>Программа не нашла на документе текстовый блок «Заказчик» и «Подрядчик», этот блок был сохранен как рисунок и недоступен для поиска по PDF/A.</p> | <p>Программа нашла на документе текстовый блок «Заказчик» и «Подрядчик», эта информация стала доступна для поиска по PDF/A. Слово «Подрядчик» было корректно выделено и распознано из под печати.</p> |

Сравнительный тест ABBYY FineReader и Cognitive ScanPack. Распознанный текст удалось выделить фиолетовым цветом. (Тест проведен и предоставлен кафедрой инженерной кибернетики НИТУ МИСиС)

Глава технологической лаборатории Владимир Арлазаров ответил на претензию Abbuu, заявив, что формат PFD/A для сжатия изображений и хранения документов в своих продуктах и технологиях действительно используют многие разработчики. При этом применяется технология MRC (Mixed Raster Content), которая является расширением подхода, используемого в формате DjVu. При использовании MRC проводится геометрическая сегментация с использованием технологий распознавания, при которой изображение расслаивается на графические слои (картинка и текст), для которых используются различные алгоритмы сжатия.

По словам Арлазарова, у этого подхода есть серьезный минус: если система не сможет распознать объект (текст на картинке, печать или подпись на печатном тексте, плохое качество ксерокопии, книгу или газету на «желтой» бумаге), то он будет обработан как изображение, что сделает невозможным поиск по нему в итоговом документе.

В Cognitive ScanPack, поясняет Арлазаров, применена цветовая и геометрическая сегментация, которая позволяет выделять в документе несколько «информационных слоев», благодаря чему способна обрабатывать текст при наложении на него печати или подписи, при обводке текста фломастером, при зачеркивании или при сильных «шумах» из-за фона бумаги, артефактов ксерокопирования или жирных пятен.

Разбивка документа на независимые слои важна при работе с документами, в которых фон является значимым, например, при обработке паспортов.

Кроме того, говорит Арлазаров, «методы бинаризации, использованные для восстановления текста ScanPack повышают визуальное качество текста на итоговом документе по сравнению с исходным». После этого каждый информационный слой обрабатывается наиболее эффективным алгоритмом сжатия (текст сжимается в TIFF, изображения, как правило, в JPG).

Вице-президент по маркетингу Cognitive Technologies Николай Никольский утверждает, что продукты на основе ScanPack не будут прямыми конкурентами решениям Abbyy. Владимир Арлазаров добавляет, что, хотя по умолчанию в ScanPack используется ядро распознавания Cuneiform, при желании пользователь сможет подключить к системе Abbyy FineReader.

Интересно, что ScanPack, умеющий распознавать и выделять изображения печатей и подписей, способен упростить возможность фальсификации бумажных документов. Владимир Арлазаров признает, что с появлением на массовом рынке продуктов на основе ScanPack, будет облегчена подделка документов злоумышленниками. Однако, говорит он, она и сейчас вполне доступна всем желающим, освоившим Photoshop.

По словам Арлазарова, разработчики постараются снять опасность злоупотреблений своей технологией, например, путем добавления к итоговому документу водяных знаков или искусственным снижением качества воспроизведенных подписей и печатей.

Как говорят в Cognitive, технологии, на которых основан ScanPack, внедрены в страховой компании «Цюрих страхование», в Магнитогорском металлургическом комбинате и, насколько известно CNews, используется в силовых структурах (что не подтверждается и не опровергается руководством Cognitive).

Николай Никольский говорит, что решения на основе Cognitive ScanPack будут выведены на массовый рынок в течение 2011 г. Объем отечественного рынка «систем структурного анализа документов» он оценивает как \$1 млрд. Оценить глобальный спрос на такие системы он затруднился, однако предположил, что, при практическом отсутствии аналогов, Cognitive ScanPack способен занять существенную долю и на мировом рынке.

Интересно, что ScanPack в значительной степени основан на свободных технологиях: ядро распознавания Cuneiform разработано Cognitive и опубликовано в 2008 г. под свободной лицензией BSD, а PDF/A - это подмножество PDF, стандартизованное в системе ISO. Компоненты распознавания и обработки изображений, как рассказали CNews в Cognitive, напротив, сейчас находятся в процессе патентования.