УДК 004.932.2

А.А. Аллаберганов, М.Ю. Катаев

Методика получения текстовой информации из изображений и ее анализ (многофункциональный исследовательский комплекс)

Текстовая информация, представленная на бумаге (бумажный носитель), часто переносится в цифровой вид как изображение определенного формата и значит может быть помещена в электронный файл, например формата «PDF». Если для анализа используется непосредственно бумажный документ, то могут применяться одни методы выделения текста, в случае цифровой формы используются методы цифровой обработки изображений. Проблема в том, что число методов обработки изображений достаточно велико и их применение для каждого конкретного случая требует соответствующих обоснований. Для выделения элементов текста, представленного в электронном формате (файле), и распознания текстовой информации (характеристики – шрифт, чернила, оттиск печати и др.) необходимо использовать специально разработанные подходы. В работе приведена методика получения текстовой информации из изображений в целях криминалистики.

Ключевые слова: изображение, текст, методики анализа, распознание, обработка изображений.

Целью работы является попытка автоматизации процессов измерения, анализа и сопоставления текстовой информации для решения задач экспертизы в исследовательской криминалистической деятельности. Это позволит обеспечить данное направление новым видом решений и качеством идентификации объектов исследования. Представлено описание измерительной (исследовательской) установки и основные ее показатели. Показан пример решения криминалистической экспертизы при анализе документа в электронном формате (файл).

В данной работе предлагается решение проблемы извлечения текстовой информации из изображения документа. Проблема приобрела множество интересных решений благодаря практическим приложениям, таким как распознавание текстов, перевод, криминалистика и др. Несмотря на то, что этой проблемой занимаются много времени, остаются области исследований, связанные с распознаванием текстовой информации документов. Качество текстовой информации документов, преобразованной в изображениях, зависит от типа сцены (простая или сложная), типа цифровой камеры (недорогие устройства, мобильные устройства и др.), величины и направления освещенности, типа подложки (бумага, пластик или др.) и т.д. Указанные моменты приводят к тому, что автоматическое извлечение текста из изображения чрезвычайно сложно.

Текст на изображениях содержит значимую и полезную информацию для понимания содержания изображений, что играет важную роль в анализе самих документов. Изображение документа содержит различную информацию, такую как тексты, рисунки и графики, что представляет собой сплошные или штриховые линии. Сложность извлечения этих линий (составляющих текст) состоит в изменении качества изображения путем сканирования, длительной истории самого документа или его изображения и др. В случае цветного изображения документа извлечение текста становится сложным и не всегда возможно различать отдельные составляющие текста (линии, составляющие буквы) из-за смешивания цветов текста и фона. Методы обнаружения текста можно классифицировать на три категории.

Первая состоит из связанных между собой методов, которые предполагают, что области изображения, содержащие текст, должны иметь однородные цвета, удовлетворять определенным размерам, заданной форме и т.д. Эти методы эффективны лишь только в случае высокого контраста между цветом текста и фона, а в других случаях эффективность низкая.

Вторая – из методов, вычисляющих текстурные показатели, которые предполагают, что области изображения, содержащие текст, имеют текстуру, отличную от фона. Хотя эти методы сравнительно менее чувствительны к цвету фона, они могут не различать текст от текстоподобного фона.

Третья – из методов, основанных на вычислении линий на изображении и поиска связности между ними. Области текста обнаруживаются в предположении, что краевые перепады фона меньше, чем у текстовых областей. Однако такого рода подходы не очень эффективны для обнаружения текстов с большими размерами шрифта.

Изображения документов отличаются от других изображений (город, природа, портрет и др.), поскольку они содержат в основном текст и, как правило, однородный фон. Фон может быть неоднородным, но его структура является типичной и повторяющейся по документу. Проблемной стороной являются изображения, преобразованные в программах электронных документов, таких как Acrobat Reader (PDF) или PowerPoint (PPT) и т.д.

Быстрое развитие цифровых технологий привело к цифровой форме представления всех категорий документов и иных текстовых материалов (документы, статьи, книги и др.) в виде изображений.

Входной информацией в системах электронного документооборота, экспертизы, контроля и других приложений могут быть не только документы с печатным текстом, но и рукописные документы.

Задача получения изображений и распознавания текста (РТ) известна давно, но до сих пор имеются как теоретические, так и практические проблемы,

связанные с огромным многообразием языков и типов написания символов и текста. Для некоторых известных методик входными данными являются изображения, полученные с разных цифровых устройств.

В свою очередь, предлагаемая нами методика исследования текстовой информации, представленной в виде изображения и распознавания фальсификации (подделки) документа в электронном формате (машинописного или рукописного) текста, оттисков печатей, является актуальной и востребованной в различных сферах практической деятельности.

Постановка задачи

Предметом исследования в данной работе является криминалистический анализ документов, предметов археологии или искусства и т.д. Цель исследований связана с повышением точности идентификации изучаемых объектов. Алгоритм определения и распознавания фальсификации (подделки) документа в электронном формате представлен на рис. 1.

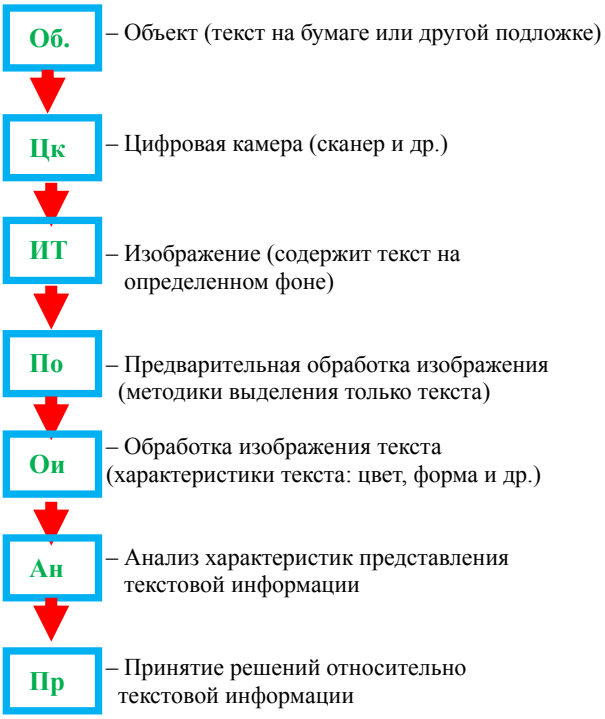


Рис. 1. Алгоритм предлагаемой методики обработки изображений с целью выделения текстовой информации

Изображение текстовой информации может быть монохромным (Gray), бинарным или цветным (RGB). Эти особенности позволяют выделять на изображении текстовую информацию при помощи соответствующих математических алгоритмов. Выделение шрифта, типа чернил связано с задачей определения лица, написавшего текст (напечатавшего), времени печати, места печати и др.

Текстовая информация содержит 5 элементов: фон бумаги; цвет чернил; текст (рукописный или машинописный); тип прибора для машинописного текста (принтер, сканер и др.) и объекты (например, оттиск печати). При обработке изображений, содержащих текст, получаем информацию в каналах {R, B, G}. Используя стандартные библиотеки обработки изображений, можно оценить качество изображения, даваемого оптической системой (цифровой камерой). Для очистки изображения от шумовой составляющей применяется далее цифровая фильтрация изображений.

Обработка и анализ могут быть проведены не обязательно в области измеряемых значений, а в области, например, спектрального пространства (например, собственных векторов, вэйвлет-преобразований дискретного косинусного преобразования и др.). Это позволяет определить признаки объектов, которые присутствуют на исследуемом (исходном) изображении.

Полученные результаты

В качестве примера приведем исследование и распознавание фальсификации (подделки) документа в электронном формате «изображение» файле «PDF» (машинописного и рукописного) текста, а также оттиска печати и подписи.

Для примера работы предлагаемой нами методики выбран документ всемирно известной американской корпорации, на котором есть оттиск печати и подпись, также в документе имеются машинописный и рукописный тексты. Для решения задач изменения изображений нами используется многофункциональная установка (КМК), подробно описанная в работе [4]. В результате обработки изображения документа необходимо установить: а) как и каким способом был изготовлен данный документ; б) применяемые технические средства и приемы при изготовлении данного документа.

Данные вопросы выясняется (исследуются) с помощью КМК.

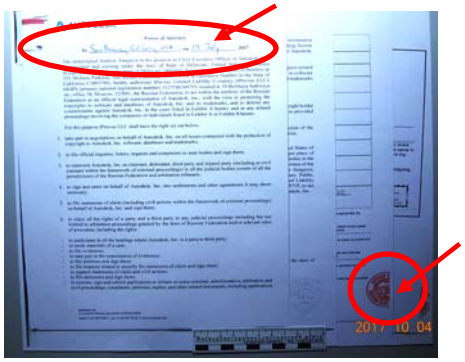


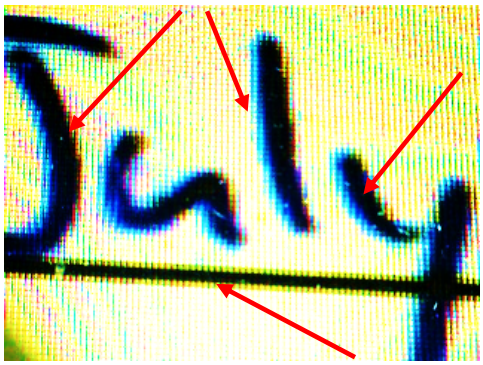
Рис. 2. Пример изображения текстовой информации, имеющей плохое качество печати

В предлагаемой вниманию задаче имеется документ очень плохого качества изготовления, с которого было получено изображение, которое перемещено в файл PDF. Доступ к самому документу закрыт, и необходимо только на основе обработке изображения данного документа получить ответ на ранее заданные вопросы.

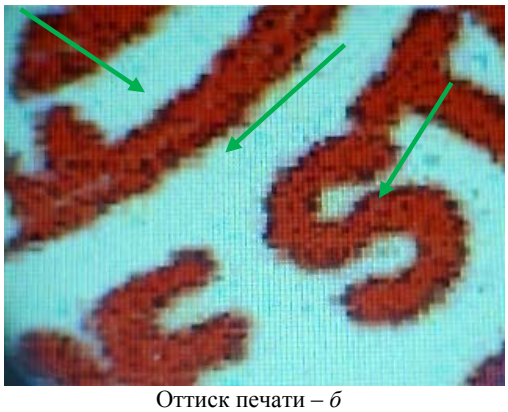
Получение результата – выявление подделки документа – представлено на рис. 3.

Заметим, что получение обычных изображений не всегда позволяет получить ожидаемый результат, однако является обязательным первым шагом. Предварительная обработка изображения и анализ границ

текстовой информации, выявленной, например, методом Хафа, позволяет выявить текст. Особенности нанесения текстовой информации возможно подчеркнуть методом увеличения пространственного разрешения при помощи специализированного оборудования (видеомикроскопа). Именно эта информация и показана на рис. 3.



Рукописный текст – *а*



Оттиск печати – *б*

Рис. 3. Исследование текстовой информации изображения с применением специализированных видеомикроскопов

Выделение букв и оценка ширины границ позволяет получить с определенной точностью, оценку метода нанесения текста на бумагу. Понятно, что механический способ нанесения печати будет отличаться от нанесения печати лазерным или струйным печатающим устройством. Последний способ приводит к размытию границы текста и брызгам чернил, которые производит форсунка. Эти все элементы могут быть получены методиками технического зрения.

Анализируя изображение рис. 3, на основе результатов обработки можно сказать, что рукописный текст (см. рис. 3, *а*) и оттиск печати (см. рис. 3, *б*) изготовлены при помощи цветного струйного принтера. Текст и печать изготовлены с применением технических средств и приемов (техническая подделка документов).

Заключение

Предлагаемая методика является первым шагом к полной автоматизации процесса извлечения текстовой информации из документов и ее анализу. Первый шаг возникает за счет того, что человек-эксперт оценивает результаты извлечения текста из изображения и принимает соответствующее решение. Однако чтобы принять правильное решение, извлекается информация по нескольким направлениям: цветовая структура изображения самого текста и окружающего пространства, форма текста, расстояние между элементами текста и др. Второй шаг будет сделан после накопления статистической информации о вариациях изменений извлеченной текстовой информации шаблонов с известными характеристиками.

Литература

1. Фомин Я.А. Распознавание образов: теория и практика. – 3-е изд., доп. – М.: ФАЗИС, 2014. – 460 с.

2. Журавлев Ю.И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлев, В.В. Рязанов, О.В. Сенько. – М.: ФАЗИС, 2006. – 176 с.

3. Местецкий Л.М. Математические методы распознавания образов. – М.: МГУ, ВМиК, 2002. – 85 с.

4. Аллаберганов А.А., Катаев М.Ю. Многофункциональный исследовательский комплекс решения задач анализа текстовой информации: матер. конф. «ЭСиСУ»: в 2 ч. – Томск: В-Спектр, 2018. – Ч. 1. – 227 с.

Аллаберганов Ахмеджан Атаханович
Аспирант каф. АСУ ТУСУРа
Эл. почта: nsk-kapital@mail.ru

Катаев Михаил Юрьевич
Д-р техн. наук, профессор каф. АСУ ТУСУРа
Эл. почта: kmy@asu.tusur.ru