УДК 004.3

Д.А. Сысоева
Донецкий национальный технический университет, г. Донецк
кафедра прикладной математики и информатики

СОВРЕМЕННЫЕ ПОДХОДЫ К ПОИСКУ ИЗОБРАЖЕНИЙ, СОДЕРЖАЩИХ ТЕКСТ

Аннотация

Сысоева Д.А. Современные подходы к поиску изображений, содержащих текст. *В работе рассмотрены комбинированные методы обнаружения текста на графических изображениях. Выделены основные способы обнаружения текста на изображениях формата «документ» и произвольных изображениях. Проведена оценка методов на основе выделенных проблем и основных характеристик алгоритма.*

Ключевые слова: *комбинированные методы, обнаружение текста, произвольные изображения, оценка методов.*

Постановка проблемы. Спам в электронной почте является самым популярным способом распространения нежелательной информации. В ответ на разработку спам-фильтров, авторы нежелательных сообщений стали рассылать тексты в виде изображений, прикрепленных к письмам. В связи с этим современные спам - фильтры и антивирусное программное обеспечение должны использовать технологии, позволяющие обнаруживать текстовый спам, внедренный в изображение. Для решения этой задачи изначально идентифицировали наличие текста в изображении, а затем определяли, является ли данный текст спамом. На данный момент многие лаборатории по защите от спама ставят акцент на этап обнаружения текста, оставляя получателю право окончательного решения по поводу нежелательности письма, содержащего такие вложения.

В информационно-поисковых системах одним из способов поиска является поиск по информации, содержащейся в самом изображении. Источниками картинок являются базы торговых знаков, фотостоки, сети Интернет, где изображения могут обладать произвольным содержанием. Для создания системы всестороннего содержательного поиска в электронных коллекциях изображений необходимо совершенствовать методы обнаружения текста.

Одним из этапов распознавания образов является выявление или обнаружение текста на изображении.

Цель статьи состоит в исследовании сложившихся в настоящее время подходов к обнаружению текста в изображениях, выявлении их ограничений и недостатков.

Постановка задачи исследования. Задача выявления или обнаружения текста в изображении решается путем анализа содержимого изображения, вычисления характеристик его содержимого, на основе которых изображении может быть отнесено к группе интереса. При этом каждое изображение Pk, k=1, 2..V, представляет собой матрицу M*N, элементы которой содержат цвета соответствующих пикселей изображения; M и N - ширина и высота изображения, а содержимое каждого изображения Pk характеризуется некоторой метамерой Fk, k=1,2..V.

Существующие методы и подходы. В настоящее время интерес к проблеме выявления текста в изображениях проявляют ряд исследователей в разных странах. Целью данной статьи является изучение существующих задач, связанных с поиском изображений, содержащих текст, и существующих методов и подходов к их решению.

Очевидно, что использование традиционных методов, которые используются для сопоставления содержимого изображений, например, методов, ориентированных на использование гистограммных признаков, точечных оценок, кластеризации и сегментации, не решит рассматриваемую проблему из- за специфики рассматриваемого класса изображений. Однако традиционные методы могут быть использованы как основа, в качестве отдельных этапов при решении рассматриваемой задачи.

Задача обнаружения текста на изображениях включает в себя такие подзадачи: выделение возможных текстовых областей, определение угла поворота текста и определение порядка чтения для определения логической структуры страницы. Для выделения текстовых областей существуют две группы методов: методы, использующие гистограммы, и методы, использующие сегментацию.

Гистограммы используются в основном для выделения текста на изображениях формата «документ». На таких картинках текст структурирован, выровнен по странице и горизонтален, имеет схожий шрифт и, если присутствует наклон, то одинаковый для всех строк. Существует два способа создания гистограмм: сверху – вниз, в котором при поиске оценивается вся страница целиком, и снизу – вверх, где для определения текстовых областей выделяются компоненты связности - символы.

К первому способу относится метод проекций страницы (Projection profiles and XY cuts) [1]. Основной идеей метода является подсчет черных пикселей в выделенной строке и создание гистограммы вертикальной или горизонтальной проекций. Затем происходит «разрезание» областей по светлым долинам гистограммы. Горизонтальные и вертикальные проекции могут накладываться друг на друга. После построения проекции страницы строится проекция для строки, где строятся проекции для отдельных слов. Метод дает хорошие результаты, если исходный документ не содержит изображений.

Представителями метода «снизу - вверх» являются алгоритм Docstrum и алгоритм с использованием диаграмм Вороного [1]. Гистограммы для этих алгоритмов строятся на основании расстояния между компонентами связности: символами, словами, строками, разделами.

Для выделения текста на произвольных изображениях существует большее количество методов, чем для изображений типа документ. Их можно разделить на методы, использующие текстуры (texture - based) и методы использующие области выделения (region - based) [1], [9], [10].

Текстура текста явно отличается от текстуры обычного изображения. Для определения признаков текстуры строится «пирамида» изображений, уменьшенных или увеличенных по размеру. Затем происходит проход по пикселям всей группы изображений и с помощью косинус- или вейвлет-преобразований определяются признаки текстуры. Этот метод хорош для однонаправленных текстов с одинаковым размером шрифта.

Методы выделения краев объектов (Edge detection), анализа связных компонент (CCA - Connected Component Analysis), выделения углов (Corner detection), использование постоянства ширины штриха (Stroke Width Transform) используют большое количество эвристик, помимо основных своих показателей. Данные алгоритмы относительно просты в реализации, и справляются с произвольным направлением и шрифтом текста [1], [3], [8].

Для нахождения угла и поворота текста используют нахождение центра масс, выделение краев с помощью матричных фильтров Собеля, Хафа, Кэнни [1], [2].

Описание и анализ исследуемых методов. Для большей вероятности и точности обнаружения текста все алгоритмы состоят из комбинаций различных методов. В данной статье рассмотрены четыре комбинированных метода, в основу которых заложены разные подходы.

В работе [5] рассматривается метод «text detection in individual video images» (далее – метод 1), предполагающий выполнение четырех этапов. Первый этап предназначен для преобразования цветного изображения в полутоновое изображение (grayscale) и дальнейшее получение бинарного образа. Он обнаруживает потенциальные регионы текста. На втором этапе эти регионы обрабатываются путем применения к ним нескольких пространственных фильтров для удаления шума и фрагментов, которые не содержат текста. Затем, на этапе распознавания обрабатывается каждое текстовое поле, причем поля, которые не прошли распознавание, удаляются.

В работе [4] рассматривается метод быстрого и эффективного выделения текста (Fast and effective text detection), в основе которого лежат два основных алгоритма: штриховой фильтр (Stroke Width Filter) и метод опорных векторов (Support vector machines). Stroke Filter - определяет потенциальные блоки текстов. SVM - определяет и извлекает строки текста (далее – метод 2).

Комбинированный метод “Fuzzy image processing” (далее – метод 3) использует три алгоритма: обнаружение, распознавание и нечеткая система (Fuzzy system). Он работает с grayscale изображениями [6].

Описанный в работе [7] метод особо контрастных пикселей (далее – метод 4) основан на подсчете количества особо контрастных соседей для каждого пикселя, на изображениях с оттенками серого цвета. Он содержит четыре этапа: конвертирование цветных изображений в оттенки серого цвета; обход восьми соседей; сегментация изображения; поиск блоков строк.

Для сравнения рассмотренных алгоритмов обнаружения текстов на изображениях были выбраны показатели, значения которых приводятся разработчиками в работах [4],[5],[6],[7]. Такими показателями являются скорость работы алгоритмов, простота программной реализации и процент обнаружения. Простота реализации характерна для методов 2 и 4, эти же методы характеризуются лучшим быстродействием. Данные о проценте распознавания приведены в таблице 1.

Таблица 1 – Процент распознавания изображений, содержащих текст, при использовании различных методов

Название	Процент обнаружения
Метод 1	55-59
Метод 2	68
Метод 3	63
Метод 4	75-80

Наибольшие проблемы с выделением текста возникают для зашумленных изображений, а также тех, в которых текст расположен под углом, невелика контрастность (цвет символов близок к цвету фона), фон является сложным либо присутствует наложение текста. В идеале распознавание изображения должно быть успешным при наличии любой из указанных проблем. Результаты анализа рассмотренных четырех методов с точки зрения перечисленных проблем приведены в таблице 2.

Таблица 2 – Решение проблем обнаружения текста при использовании различных методов

Наименование	Проблемы				
	Отделение шумов	Текст под углом	Цвет букв близок к фону	Сложный фон	Наложение текста
Метод 1	+	-	-	+	+
Метод 2	-	-	-	-	+
Метод 3	+	+	-	-	+
Метод 4	+	+	-	+	+

На основании данных разработчиков можно утверждать, что с точки зрения устойчивости к проблемам, вызывающих затруднения распознавания текста на изображениях, и по основным показателям работы алгоритма наиболее эффективным является метод особо контрастных пикселей. Данный метод реализует наилучшее обнаружение текста, который может находиться на разных типах фона.

Выводы. Выполненный в работе анализ различных методов и подходов к обнаружению текста на изображении показал, что существующие методы недостаточно эффективны при решении рассматриваемой задачи, поскольку не позволяют преодолеть все проблемы, вызывающие затруднения распознавания текста на изображениях. Показано, что по основным показателям работы наиболее эффективным является метод особо контрастных пикселей.

Список литературы

1. Анализ изображений и видео, лекция 9/ Интернет-ресурс. - Режим доступа: www/ URL: http://rudocs.exdat.com/docs/index-87395.html
2. Д.А. Форсайт, Ж. Понс Компьютерное зрение. Современный подход: Пер. с англ. – М.: Издательский дом «Вильямс», 2004. – 928 с.
3. Ali Mosleh, Nizar Bouguila, A. Ben Hamza, « Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform», In Proceedings British Machine Vision Conference 2012. Pages 63.1- 63.12.
4. Xiaojun Li, Weiqiang Wang, Shuqiang Jiang, Qingming Huang, Wen Gao, «Fast and effective text detection» 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, 12-15 Oct. 2008, P. 969 – 972
5. Chein-I Chang, Yingzi Du, «Automated system for text detection in individual video images», Journal of Electronic Imaging / July 2003 / Vol. 12(3)
6. Mohanad Alata — Mohammad Al-Shabi, «Text detection and character recognition using fuzzy image system», Journal of electrical engineering, Vol. 57, NO. 5, 2006, 258–267
7. Обнаружение текста/ Интернет-ресурс. - Режим доступа: www/ URL: http://macarov.net/news/obnaruzhenie_teksta/2010-03-03-13
8. Методы обнаружения текста Интернет-ресурс. - Режим доступа: www/ URL: http://rudocs.exdat.com/docs/index-78161.html
9. Обнаружение текста на изображениях / Интернет-ресурс. - Режим доступа: www/ URL: http://www.lektorium.tv/lecture/?id=13721
10. Задача распознавания изображений / Интернет-ресурс. - Режим доступа: www/URL:http://gendocs.ru/v2103