Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

А.В. Бондаренко, В.А. Галактионов, В.И. Горемычкин,
А.В. Ермаков, С.Ю. Желтов

Исследование подходов к построению систем автоматического считывания символьной информации.

Москва 2003

УДК 519.68

А.В.Бондаренко, В.А.Галактионов, В.И.Горемычкин, А.В.Ермаков, С.Ю.Желтов **Исследование подходов к построению систем автоматического считывания символьной информации.**

Данная работа посвящена разработке прототипа алгоритмического и программного обеспечения для информационных систем автоматизации документооборота. Рассмотрены основные принципы построения системы автоматического считывания. Предложены оригинальные алгоритмы обнаружения текстовых строк, определения знако-мест и адаптивной бинаризации изображений. Для распознавания символов OCR-B-кода реализованы метод зон и метод пересечений.

A.V.Bondarenko, V.A.Galaktionov, V.I.Goremychkin, A.V.Yermakov, S.Y.Zgeltov **Research of the approaches to construction of systems of automatic reading of the symbolical information.**

The given work is devoted to development of the prototype of algorithm and software for information systems of automation of document circulation. The basic principles of construction of automatic reading system are considered. The original algorithms of detection of textual lines, determination of symbol places and adaptive binarization of the images are offered. Method of zones and method of crossings are implemented for recognition of OCR-B-code symbols.

1. Введение.

Задача распознавания текстовой информации при переводе печатного и рукописного текста в машинные коды является одной из важнейших составляющих проектов, имеющих целью автоматизацию документооборота. Вместе с тем эта задача является одной из наиболее сложных и наукоемких в области автоматического анализа изображений. Даже человек, читающий рукописный текст, в отрыве от контекста делает в среднем 4% ошибок. Что касается систем считывания печатных документов, то здесь сложность заключается в том, что в ответственных приложениях, таких как, например, автоматизация ввода паспортно-визовой информации, необходимо обеспечить высокую надежность распознавания (более 98-99%) даже при плохом качестве печати и оцифровки исходного текста.

В последние десятилетия, благодаря использованию современных достижений компьютерных технологий, были развиты новые методы обработки изображений и распознавания образов ([l]-[ll]), благодаря чему стало возможным создание таких систем распознавания печатного текста, которые удовлетворяли бы основным требованиям систем автоматизации документооборота. Тем не менее, создание каждого нового приложения в данной области по-прежнему остается творческой задачей и требует дополнительных исследований в связи со специфическими требованиями по разрешению, быстродействию, надежности распознавания и объему памяти, которыми характеризуется каждая конкретная задача разработки проблемно-ориентированной системы автоматического ввода в компьютер бумажной документации.

Различные технологии, объединенные под общим термином *"распознавание символов",* подразделяются на распознавание в реальном

режиме времени и распознавание в пакетном режиме, каждый из которых характеризуется собственной аппаратной частью и собственными алгоритмами распознавания.

В данной работе будет рассматриваться только распознавание в пакетном режиме, то есть такое, которое осуществляется после завершения ввода печатных символов.

В типичной системе *оптического распознавания текстов* (OCR) вводимые символы читаются и оцифровываются оптическим сканером. После этого каждый символ подвергается локализации и выделению, и получившаяся матрица подвергается предобработке, т. е. сглаживанию, фильтрации и нормализации. В результате предобработки выделяются характерные признаки, после чего производится классификация. В работе описываются базовые идеи и методы, позволяющие решить указанные подзадачи, а также их модификации, используемые в современных системах распознавания.

2. Типовые проблемы, связанные с распознаванием символов

Существует ряд существенных проблем, связанных с распознаванием рукописных и печатных символов. Наиболее важные из них следующие:

- разнообразие форм начертания символов;
- искажения изображений;
- вариации размеров и масштаба символов.

Каждый отдельный символ может быть написан различными стандартными шрифтами, например (Gothic, Elite, Courier, Orator), специальными шрифтами, использующимися в системах OCR, а также множеством нестандартных шрифтов. Кроме того, различные символы могут обладать сходными очертаниями. Например, 'U и 'V, 'S' и '5', 'Z' и '2', 'G' и '6'.

Искажения цифровых изображений символов могут быть следующих

видов:

Искажения формы: разорванность строк, непропечатанность символов, изолированность отдельных точек, неплоский характер информационного носителя (например, эффект коробления), смещения символов или их частей относительно местоположения в строке;

вращения с изменением наклона символов;

грубым дискретом оцифровке изображений;

Кроме того, необходимо выделить радиометрические искажения: дефекты освещения, тени, блики, неравномерный фон, ошибки при сканировании или при съемке видеокамерой.

Существенным является и влияние исходного масштаба печати. В принятой терминологии масштаб 10, 12 или 17 означает, что в дюйме строки помещаются 10, 12 или 17 символов. При этом, например, символы масштаба 10 обычно крупнее и шире символа масштаба 12.

Помимо указанных проблем, система оптического распознавания текста (OCR), должна выделять на изображении текстовые области, в них выделять отдельные символы, распознавать эти символы и быть нечувствительной к способу печати (верстки) и расстоянию между строками.

3. Структура систем оптического распознавания текстов.

Как правило, системы OCR состоят из нескольких блоков, предполагающих аппаратную или программную реализацию:

- оптический сканер;
- блок локализации и выделения элементов текста;
- блок предобработки изображений;
- блок выделения признаков;
- блок распознавания;
- блок постобработки результатов распознавания.

В результате работы оптического сканера исходный текст вводится в компьютер в виде полутонового или бинарного изображения.

В целях экономии памяти и уменьшения затрат времени на обработку информации, в системах OCR, как правило, применяется преобразование полутонового изображения в черно-белое. Такую операцию называют *бинаризацией.* Однако необходимо иметь в виду, что операция бинаризации может привести к ухудшению эффективности распознавания.

Программное обеспечение в системах OCR отвечает за представление данных в цифровом виде и разбиение связного текста на отдельные символы.

После разбиения символы, представленные в виде бинарных матриц, подвергаются сглаживанию, фильтрации с целью устранения шумов, нормализации размера, а также другим преобразованиям с целью выделения признаков, используемых впоследствии для распознавания.

Распознавание символов происходит в процессе сравнения выделенных характерных признаков с эталонными признаками, отбираемыми в ходе статистического анализа результатов, полученных в процессе обучения системы.

Таким образом, смысловая или контекстная информация может быть использована как для разрешения неопределенностей, возникающих при распознавании символов, обладающих идентичными размерами, так и для корректировки слов и фраз в целом.

4. Оптическое сканирование изображений.

Одна из наиболее ранних попыток создать систему, способную считывать тексты, была предпринята в 1870 году. Она представляла собой сканер-сетчатку, работа которого была основана на фотоэлементах. В дальнейшем появились Fourier d'Albe's Optophone в 1912 г. и Thomas tactile relief device в 1926 г. Системы оптического считывания текстов появились в

середине XXв. в результате развития цифровых компьютеров. Дэвид Шепард, основатель компании Intelligent Machine Research, считается родоначальником создания коммерческих систем OCR. Несмотря на возможность считывания весьма ограниченного числа шрифтов и на ограничения, накладываемые на ориентацию символов, в начале 60-х г.г. системы автоматического считывания получили широкое распространение. С развитием микроэлектроники эти системы постоянно совершенствовались.

В настоящее время наиболее распространены следующие методы сканирования:

- Раздельно-щелевое сканирование. Этот метод использует массив фотоэлектрических элементов.
- Сканирование лазерным лучом.
- Сканирование матрицей фотоэлементов.
- Механическое сканирование (диском Нипкова). Принцип действия основан на вращении диска с щелями, сквозь которые проходящий отраженный свет попадает на фотодиоды.

В прошлом раздельно-щелевое сканирование активно применялось в системах OCR. Данный метод позволял производить сканирование по горизонтали, тогда как сканирование по вертикали обеспечивалось перемещением сканируемого документа.

Напечатанный документ перемещается по освещенной области. Отраженный свет, собранный линзами, попадает на фотодиоды, расположенные по горизонтали. Блок видеоусиления увеличивает амплитуду сигналов, поступающих от фотодиодов, и преобразует их в сетку черных и белых точек. Раздельно-щелевое сканирование последнего поколения использует технологию *приборов с зарядовой связью* (ПЗС матрицы).

5. Методы предобработки изображений текстовых символов.

Предобработка является важным этапом в процессе распознавания

образов и позволяет производить сглаживание, нормализацию, сегментацию и аппроксимацию отрезков линий.

Сглаживание состоит из операций *заполнения* и *утоньшения. Заполнение* устраняет небольшие разрывы и пробелы.

Утоньшение представляет собой процесс уменьшения толщины линии, в которой сразу несколько пикселов ставятся в соответствие только одному пикселу. Известны последовательные, параллельные и гибридные алгоритмы утоньшения. Наиболее общие методы утоньшения основаны на итеративном размывании контуров, при котором окно (3x3) движется по изображению, и внутри окна выполняются соответствующие операции. После завершения каждого этапа все выделенные точки удаляются.

Нормализация состоит из алгоритмов, устраняющих перекосы отдельных символов и слов, а также включает в себя процедуры, осуществляющие нормализацию символов по высоте и ширине после соответствующей их обработки.

Сегментация осуществляет разбиение изображения на отдельные области. Как правило, прежде всего необходимо очистить текст от графики и рукописных пометок, поскольку перечисленные методы позволяют обрабатывать лишь незашумленный текст. Очищенный от различных пометок текст уже может быть сегментирован.

Большинство алгоритмов оптического распознавания разделяют текст на символы и распознают их по отдельности.

Это простое решение действительно эффективно, если только символы текста не перекрывают друг друга. Слияние символов может быть вызвано типом шрифта, которым был набран текст, плохим разрешением печатающего устройства или высоким уровнем яркости, выбранном для восстановления разорванных символов.

Разбиение текста на слова возможно в том случае, если слово является состоятельным признаком, в соответствии с которым выполняется сегментация. Подобный подход сложно реализовать из-за большого числа

элементов, подлежащих распознаванию, но он может быть полезен, если набор слов в кодовом словаре ограничен по условию задачи.

Под *аппроксимацией отрезков линий* понимают составление графа описания символа в виде набора вершин и прямых ребер, которые непосредственно аппроксимируют цепочки пикселов исходного изображения. Данная аппроксимация осуществляется в целях уменьшения объема данных и может использоваться при распознавании, основанном на выделении признаков, описывающих геометрию и топологию изображения.

6. Признаки символов, используемые для автоматического распознавания.

Считается, что выделение признаков является одной из наиболее трудных и важных задач в распознавании образов. Для распознавания символов может быть введено большое количество различных систем признаков. Проблема заключается в том, чтобы выделить именно те признаки, которые позволят эффективно отличать один класс символов от всех остальных.

В данном разделе описан ряд основных методик распознавания символов.

6.1. Корреляция и сопоставление шаблонов.

Введенная матрица символов сравнивается с набором эталонов. Вычисляется степень сходства между образом и каждым из эталонов. Классификация тестируемого изображения символа происходит по методу ближайшего соседа.

С практической точки зрения этот метод легко реализовать, и многие коммерческие системы его используют. Однако даже небольшое темное

пятно, попавшее на внешний контур символа, может существенно повлиять на результат распознавания. Поэтому для достижения хорошего качества распознавания в системах, использующих сопоставление шаблонов, применяются другие способы сравнения изображений.

- Одна из основных модификаций алгоритма сравнения шаблонов использует представление шаблонов в виде набора логических правил.

6.2. Статистические распределения точек.

В данной группе методов выделение признаков осуществляется на основе анализа различных статистических распределений точек. Наиболее известные методики этой группы используют *вычисление моментов и подсчет пересечений,*

Моменты различных порядков с успехом используются в таких задачах обработки изображений как зрение роботов, обнаружение и распознавание летательных аппаратов и судов по снимкам, анализ сцен и распознавание символов. В последнем случае в качестве признаков используют значения статистических моментов совокупности "черных" точек относительно некоторого выбранного центра.

Наиболее общеупотребительными в приложениях такого рода являются построчные, центральные и нормированные моменты.

Для цифрового изображения, хранящегося в двумерном массиве, построчные моменты являются функциями координат каждой точки изображения следующего вида:

$$m_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q f(x, y) \qquad (2.1)$$

где p,q=0,l,...,∞; M и N являются размерами изображения по горизонтали и вертикали и f(x,y) является яркостью пиксела в точке (x,y) на изображении.

Центральные моменты являются функцией расстояния точки от центра

тяжести символа:

$$m_{pq} = \sum\sum (x - \bar{x})^p (y - \bar{y})^q f(x, y) \qquad (2.2)$$

где x и y с чертой - координаты центра тяжести.

Наконец, нормированные центральные моменты получаются в результате деления центральных моментов на моменты нулевого порядка.

Следует отметить, что строковые моменты, как правило, обеспечивают низкий уровень распознавания. Центральные и нормированные моменты предпочтительнее из-за их большей инвариантности к преобразованиям изображений.

В *методе пересечений* признаки формируются путем подсчета того, сколько раз произошло пересечение изображения символа с выбранными прямыми, проводимыми под определенными углами, например, 0, 45, 90, 135 градусов. Этот метод часто используется в коммерческих системах благодаря тому, что он инвариантен к дисторсии и небольшим стилистическим вариациям символов, а также обладает достаточно высокой скоростью и не требует значительных вычислительных затрат. Так, например, представлена на рынке OCR-систем автоматическое устройство, которое читает любые заглавные буквенно-цифровые символы. В качестве алгоритма распознавания используется метод пересечений.

Существует множество других методов распознавания, основанных на выделении признаков из статистического распределения точек. Например, *метод зон* предполагает разделение площади рамки, объемлющей символ, на области, и последующее использование плотностей точек в различных областях в качестве набора характерных признаков.

В методе *матриц смежности* в качестве признаков рассматриваются частоты совместной встречаемости "черных" и "белых" элементов в различных геометрических комбинациях.

6.3. Интегральные преобразования.

Среди современных технологий распознавания, основанных на преобразованиях, выделяются методы, использующие Фурье-дескрипторы символов, а также дескрипторы границ.

Преимущества методов, использующих преобразования Фурье-Меллина, связаны с тем, что они обладают инвариантностью к масштабированию, вращению и сдвигу символа. Основной недостаток этих методов заключается в нечувствительности к резким скачкам яркости на границах. В то же время, при фильтрации шума на границах это свойство может оказаться полезным.

6.4. Структурный анализ.

Структурные признаки обычно используются для выделения общей структуры образа. Они описывают геометрические и топологические свойства символа.

Одними из наиболее используемых признаков являются штрихи и пробелы, применяемые для определения следующих характерных особенностей изображения: концевых точек, пересечения отрезков, замкнутых циклов, а также их положения относительно рамки, объемлющей символ.

Пусть матрица, содержащая утоньшенный символ, разделена на ряд областей, каждой из которых присвоены буквы A, B, C и т.д. Символ рассматривается как набор штрихов. При этом штрих является прямой линией (ℓ) или кривой (c), и соединяет некоторые две точки в начертании символа. Штрих является кривой, если его точки удовлетворяют следующему выражению:

$$ABS\left|\frac{\sum_{i=1}^{n} ax_i + by_i + c / \sqrt{a^2 + b^2}}{n}\right| > 0.69, \qquad (2.3)$$

в противном случае - это прямая. В данной формуле (x_i,y_i) является точкой, принадлежащей штриху; ax+by+c=0 - уравнение прямой, проходящей через концы штриха, коэффициент 0.69 получен опытным путем.

При введенных обозначениях признаки символа могут быть записаны, например, в виде 'AℓC' и 'AcD', что означает наличие прямой, проходящей из области 'A' в область 'C', и кривой, проходящей из области 'A' в область 'D' соответственно. Достоинство структурных признаков по сравнению с другими методами определятся устойчивостью к сдвигу, масштабированию и повороту символа на небольшой угол, а также - к возможным дисторсиям и различным стилевым вариациям шрифтов.

К сожалению, задача выделения признаков данного типа находится пока в процессе исследования и не имеет еще общепризнанного решения.

7. Классификация символов.

В существующих системах OCR используются разнообразные алгоритмы *классификации,* то есть отнесения признаков к различным классам ([1]). Они существенно различаются в зависимости от выбранных признаков и от правил классификации.

Для классификации символов необходимо в первую очередь создать библиотеку эталонных векторов признаков. Для этого на стадии обучения оператор или разработчик вводит в систему OCR большое количество образцов начертания символов. Для каждого образца система выделяет признаки и сохраняет их в виде соответствующего *вектора признаков.* Набор векторов признаков, описывающих символ, называется *классом* или

кластером.

В процессе эксплуатации системы OCR может появиться необходимость расширить базу знаний. Для осуществления данной цели некоторые системы обладают возможностью дообучения в реальном режиме времени. Однако процесс обучения требует участия оператора и существенных затрат времени, хотя и проводятся исследования, направленные на автоматизацию процесса обучения, что в будущем позволит свести к минимуму участие в нем человека-оператора.

Задачей классификации является определение класса, которому принадлежит вектор признаков, полученный для данного символа.

Алгоритмы классификации основаны на определении степени близости набора признаков рассматриваемого символа каждому из классов. Правдоподобие получаемого результата зависит от выбранной метрики пространства признаков. К наиболее известным метрикам относится Евклидово расстояние:

$$D_j^E = \sqrt{\sum_{i-1}^{N} (F_{ij}^L - F_i^L)^2} \qquad (2.4)$$

где F_{ij}^L - i-й признак из j-го эталонного вектора; F_i^L - i-й признак тестируемого изображения символа.

При классификации по методу ближайшего соседа символ будет отнесен к классу, вектор признаков которого наиболее близок к вектору признаков символа. Следует учитывать, что затраты на вычисления возрастут с увеличением количества используемых признаков и классов.

Одна из методик, позволяющих улучшить метрику сходства, основана на статистическом анализе эталонного набора признаков. При этом в процессе классификации более надежным признакам отдается больший приоритет:

$$D_j^E = \sqrt{\sum_{i-1}^{N} w_i (F_{ij}^L - F_i^L)^2} \qquad (2.5)$$

где w_i, - вес i-го признака.

Другая методика классификации, требующая знания априорной информации, основана на использовании формулы Байеса. Из правила Байеса следует, что рассматриваемый вектор признаков принадлежит классу "j", если отношение правдоподобия *K* больше чем отношение априорной вероятности класса j к априорной вероятности класса L.

8. Постобработка результатов распознавания.

В высокоточных системах OCR, таких как, например, системы считывания и обработки машиночитаемых паспортно-визовых документов, качество распознавания, получаемое при распознавании отдельных символов, не считается достаточным. В таких системах необходимо использовать также контекстную информацию. Использование контекстной информации позволяет не только находить ошибки, но и исправлять их.

Существует большое количество приложений OCR, использующих глобальные и локальные позиционные диаграммы, триграммы, п-граммы, словари и различные сочетания этих методов. Рассмотрим два подхода к решению такой задачи: *словарь* и *набор бинарных матриц,* аппроксимирующих структуру словаря.

Доказано, что словарные методы являются одними из наиболее эффективных при определении и исправлении ошибок классификации отдельных символов. При этом после распознавания всех символов некоторого слова словарь просматривается в поисках этого слова, с учетом того, что оно, возможно, содержит ошибку. Если слово найдено в словаре, это не говорит об отсутствии ошибок. Ошибка может превратить одно слово, находящееся в словаре, в другое, также входящее в словарь. Такая ошибка не может быть обнаружена без использования смысловой контекстной информации, только она может подтвердить правильность написания. Если

слово в словаре отсутствует, считается, что в слове допущена ошибка. Для исправления ошибки прибегают к замене такого слова на похожее слово из словаря. исправление не производится, если в словаре найдено несколько подходящих вариантов замены. В этом случае интерфейс некоторых систем позволяет предложить пользователю различные варианты решения, например, исправить ошибку, игнорировать ее и продолжать работу или внести это слово в словарь.

Главный недостаток в использовании словаря заключается в том, что операции поиска и сравнения, применяющиеся для исправления ошибок, требуют значительных вычислительных затрат, возрастающих с увеличением объема словаря.

Некоторые разработчики с целью преодоления трудностей, связанных с использованием словаря, пытаются выделять информацию о структуре слова из самого слова. Такая информация говорит о степени правдоподобия п-граммов (например, пары и тройки букв) в тексте. N-граммы также могут быть глобально позиционированными, локально позиционированными или вообще непозиционированными.

Например, степень достоверности непозиционированной пары букв может быть представлена в виде бинарной матрицы, элемент которой равен 1, тогда и только тогда, когда соответствующая пара букв имеется в некотором слове, входящем в словарь. Позиционная бинарная диаграмма Dy является бинарной матрицей, определяющей, какая из пар букв имеет ненулевую вероятность возникновения в конкретной позиции (i,j). Набор всех позиционных диаграмм включает бинарные матрицы для каждой пары положений.

9. Заключение.

Исследование методов и программно-аппаратных систем оптического распознавания образов позволяет сформулировать следующие выводы:

1. Современное состояние технологии автоматического распознавания печатных текстов (OCR) позволяет решать задачу автоматизации ввода паспортно-визовой информации при необходимом уровне надежности.
2. При построении системы OCR, включающей оптическое устройство оцифровки изображений, блок локализации и выделения элементов текста, блок предобработки изображения; блок выделения признаков, блок распознавания символов и блок постобработки результатов распознавания, необходимо использовать методы и алгоритмы, обладающие высокой робастностью к яркостно-геометрическим искажениям и сложным текстурным фонам.
4. В качестве таких методов и алгоритмов, могут быть использованы: процедуры определения строк, знакомест на основе модификаций преобразования Hough; методы, основанные на исследовании устойчивых статистических распределений точек; методы, использующие интегральные преобразования, а также структурный анализ символов.
4. При разработке ответственных систем OCR качество распознавания, получаемое при распознавании отдельных символов, не является достаточным. В таких системах необходимо учитывать контекстную информацию. Использование контекстной информации позволяет не только находить ошибки, но и исправлять их.

Список использованных источников

1. Дуда Р., Харт П. Распознавание образов и анализ сцен. - М.: Мир, 1986.
2. Бутаков А., Островский В. И., Фадеев И.Л. "Обработка изображений на ЭВМ", - М.: Радио и связь, 1987.
3. Andrews, H.C., Tescher, A.G., and Kruger, R.P. "Image Processing by Digital Computer.® IEEE Spectrum, vol. 9, no. 7, pp. 20-32. 1972.
4. Castleman, K.R. Digital linage Processing, Prentice-Hall, Englewood Cliffs. 1979
5. Jain, A.K. Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, N.J. 1989.
6. Fu, K.S., and Rosenfeld, A. "Pattern Recognition and Image Processing." IEEE Trans. Computers, vol. C-25, no. 12, pp. 1336-1346, 1976.
7. Gonzalez, R.C., Woods, R.E., and Swain, W.T. "Digital Image Processing: An Introduction." Digital Design, vol. 16, no. 4, pp. 15-20, 1986.
8. Pratt, W.K. Digital Image Processing, John Wiley & Sons, New York. 1978
9. Pratt, W.K. Digital Image Processing, 2nd eel., John Wiley & Sons, New York. 1991.
10. Rosenfeld, A., and Kak, A.C. Digital Picture Processing, 2nd ed., Academic Press, New York. 1982.
11. Schalkoff, R.J. Digital Image Processing and Computer Vision, John Wiley & Sons, New York. 1989.