Системы оптического распознавания документов

Системы оптического распознавания символов. Системы оптического распознавания символов используются при создании электронных библиотек и архивов путем перевода книг и документов в цифровой компьютерный формат.

Сначала с помощью сканера необходимо получить изображение страницы текста в графическом формате. Далее для получения документа в текстовом формате необходимо провести распознавание текста, т. е. преобразовать элементы графического изображения в последовательность текстовых символов.

Системы оптического распознавания символов сначала определяют структуру размещения текста на странице и разбивают его на отдельные области: колонки, таблицы, изображения и т. д. Далее выделенные текстовые фрагменты графического изображения страницы разделяются на изображения отдельных символов.

Для отсканированных документов типографского качества (достаточно крупный шрифт, отсутствие плохо напечатанных символов или исправлений) распознавание символов проводится путем их сравнения с растровыми шаблонами.

Растровое изображение каждого символа последовательно накладывается на растровые шаблоны символов, хранящиеся в памяти системы оптического распознавания.

Результатом распознавания является символ, шаблон которого в наибольшей степени совпадает с изображением

Рис. 3.16. Распознаваемый символ «Б» накладывается на растровые шаблоны символов (А, Б, В и т. д.)

Рис. 3.17. Распознаваемый символ «Б» накладывается на векторные шаблоны символов (А, Б, В и т. д.)

При распознавании документов с низким качеством печати (машинописный текст, факс и т. д.) используется векторный метод распознавания символов. В распознаваемом изображении символа выделяются геометрические примитивы (отрезки, окружности и др.) и сравниваются с векторными шаблонами символов. В результате выбирается тот символ, для которого совокупность всех геометрических примитивов и их расположение больше всего соответствует распознаваемому символу (рис. 3.17).

Системы оптического распознавания символов являются «самообучающимися» (для каждого конкретного документа они создают соответствующий набор шаблонов символов), и поэтому скорость и качество распознавания многостраничного документа постепенно возрастают.

С появлением первого карманного компьютера Newton фирмы Apple в 1990 году начали создаваться системы распознавания рукописного текста. Такие системы преобразуют текст, написанный на экране карманного компьютера специальной ручкой, в текстовый компьютерный документ.

Системы оптического распознавания форм. При заполнении документов большим количеством людей (например, при сдаче выпускником школы единого государственного
экзамена (ЕГЭ)) используются бланки с пустыми полями. Данные вводятся в поля печатными буквами от руки. Затем эти данные распознаются с помощью систем оптического распознавания форм и вносятся в компьютерные базы данных.
Сложность состоит в том, что необходимо распознавать символы, написанные от руки, которые довольно сильно различаются у разных людей. Кроме того, такие системы должны уметь определять, к какому полю относится распознаваемый текст.

Более высокие показатели могут быть достигнуты только с использованием контекстной и грамматической информации. Например, в процессе распознания искать целые слова в словаре легче, чем пытаться проанализировать отдельные символы из текста. Знание грамматики языка может также помочь определить, является ли слово глаголом или существительным. Формы отдельных рукописных символов иногда могут не содержать достаточно информации, чтобы точно (более 98 %) распознать весь рукописный текст.
Для решения более сложных проблем в сфере распознавания используются как правило интеллектуальные системы распознавания, такие как искусственные нейронные сети.

На изображениях с рукописным «печатным» текстом без артефактов может быть достигнута точность в 80 % — 90 %, но с такой точностью изображение будет преобразовано с десятками ошибок на странице. Такая технология может быть полезна лишь в очень ограниченном числе приложений.

Он-лайн системы для распознавания рукописного текста «на лету» в последнее время стали широко известны в качестве коммерческих продуктов. Алгоритмы таких устройств используют тот факт, что порядок, скорость и направление отдельных участков линий ввода известны. Кроме того, пользователь научится использовать только конкретные формы письма. Эти методы не могут быть использованы в программном обеспечении, которое использует сканированные бумажные документы, поэтому проблема распознавания рукописного «печатного» текста по-прежнему остается открытой.