Р. Ю. Готовко

ПРИМЕНЕНИЕ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ В ДОКУМЕНТООБОРОТЕ

В данной работе рассматривается применение технологии оптического распознавания символов в документообороте, типовые проблемы технологии OCR, а так же один из возможных методов постобработки результатов распознавания.

Введение

Переход от бумажного документооборота к электронному является одним из этапов развития организации. Ввиду того, что электронный документооборот локализован в рамках одной организации или отдела, то возникают ситуации, когда всё-таки приходится работать с документами на бумажных носителях, обработка которых неавтоматизированными средствами занимает продолжительное время. Автоматизировать получение текстовых данных с бумажного носителя позволяет технология оптического распознавания символов (англ. optical character recognition, OCR).

I. Типовые проблемы, связанные с распознаванием символов

Внедрение оптического распознавания символов в документооборот безусловно является одним из шагов успеха в оптимизации и автоматизации, но за данным шагом скрывается ряд существенных проблем распознавания. Наиболее важные из них:

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

Каждый отдельный символ может быть написан различными стандартными шрифтами, например (Times, Gothic, Elite, Courier, Orator), а также - множеством нестандартных шрифтов, используемых в различных предметных областях. При этом различные символы могут обладать сходными очертаниями. Например, «U» и «V», «S» и «5», «Z» и «2», «G» и «6» [1].

Искажения цифровых изображений текстовых символов могут быть вызваны:

- шумами печати, в частности, непропечаткой, «слипанием» соседних символов, пятнами и точками вблизи символов и т. п.;
- смещением символов относительно их ожидаемого положения в строке;
- изменением наклона символов;
- искажением формы символа за счет оцифровки изображения с «грубым» дискретом;
- эффектами освещения при сканировании документа.

II. Постобработка результатов распознавания

В системах OCR 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. Доказано, что словарные методы являются одними из наиболее эффективных при определении и исправлении ошибок классификации отдельных символов.

После распознавания всех символов некоторого слова словарь просматривается в поисках этого слова, с учетом того, что оно, возможно, содержит ошибку. Если слово в словаре отсутствует, считается, что в слове допущена ошибка распознавания. Для исправления ошибки прибегают к замене такого слова на наиболее похожее слово из словаря. Исправление не производится, если в словаре найдено несколько подходящих кандидатур для замены. В этом случае интерфейс некоторых систем позволяет показать слово пользователю и предложить 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.

III. Выводы

Технология оптического распознавания символов позволяет сократить временные и трудовые издержки при переносе текста с бумажного носителя на электронный, что является одним из ключевых признаков для внедрения в документооборот организации.

1. Оптическое распознавание символов (OCR) [Электронный ресурс]. – Режим доступа : http://wiki.technicalvision.ru

Готовко Роман Юрьевич, магистрант 1 курса кафедры информационных технологий автоматизированных систем БГУИР, gotovkoromanhzby@gmail.com.

Научный руководитель: Севернёв Александр Михайлович, доцент, к.т.н.