

*Муромский институт (филиал) Владимирского государственного университета
Россия, 602264, Владимирская область, г. Муром, ул. Орловская, д.23
E-mail: kanunovae@list.ru*

Применение алгоритмов распознавания для обработки скорописных текстов

Распознавание текста является одним из направлений распознавания образов [1]. Распознавание образов представляет собой очень сложную задачу в теоретическом и практическом смыслах, несмотря на то, что с ней достаточно легко справляются многие живые организмы и человек. Крайне сложно создать искусственную систему и ее технически реализовать для того, чтобы эффективно выполнять данный процесс. В данном случае, под распознаванием понимается соотнесение изображения объекта, его образа, набора признаков самому объекту.

Примерами и приложениями систем распознавания образов могут являться как распознавание текста в общем, так и отдельных его символов, распознавание речи, человеческих лиц, биометрических данных человека, штрих кодов продуктов, номеров машин и т.д.

Примерами распознавания текста являются: оцифровка изображений текста (сканированные книги, статьи, журналы) для последующей работы с его цифровым аналогом, обработка анкетных бланков, распознавание номеров машин и надписей на объектах и т.д. Задача распознавания текста остается актуальной на сегодняшний день, так как не существует сто процентной универсальной системы по распознаванию текста [2].

Распознавание текста включает в себя следующие подзадачи и подпроцессы.

1. Поступающее на вход системы изображение должно быть очищено от шума и приведено к виду, позволяющему эффективно выделять символы и распознавать их.

2. Система должна разбить изображение на блоки текста, основываясь на особенностях его выравнивания и распределения по нескольким колонкам.

3. Изображение с текстом должно быть разделено на изображения строк, а затем на изображения символов для того, чтобы в дальнейшем обработать каждый символ по отдельности. После данного шага разные системы распознавания работают по своим специфическим алгоритмам.

4. Изображение символа может обрабатываться целиком, для этого оно сравнивается с имеющимися шаблонами. Другим вариантом является выделение характеристик символа: отбор характерных признаков, и классификация данных признаков по имеющимся в системе критериям. На выходе четвертого шага появляется возможный вариант буквы. Однако обычно системы на этом не останавливаются и продолжают работу на основе других методов, уточняя полученный результат.

5. Результат распознавания может быть не удовлетворительным. Для получения более хороших результатов в системе может быть встроен блок обучения. С помощью этого блока можно задать системе примеры начертания разных букв в данном шрифте. После процесса обучения предполагается лучшее качество распознавания текста. Система распознавания текста не всегда должна следовать всем описанным шагам, но основные действия процесса распознавания являются общими для любого алгоритма.

В докладе рассматривается несколько алгоритмов распознавания образов, которых можно использовать для распознавания текста на скорописных текстах:

- распознавание по шаблонам;
- структурный подход;
- контекстное распознавание;
- нейронные сети;
- распознавание образов на базе клеточных автоматов;
- Байесовский подход для распознавания объектов;
- метод потенциальных функций;
- и другие алгоритмы.

Секция 5. Информационные технологии в образовании и производстве

Литература

1. Нейман Дж. Теория самовоспроизводящихся автоматов. Дж. Нейман. М.: Мир, 1971.
2. Wolfram S. A New Kind of Science. Wolfram Media. Inc., 2002.
3. Хайкин С. Нейронные сети: полный курс. М. : Вильямс, 2006.