УДК 004.93'14

РАСПРЕДЕЛЕННАЯ ПРОГРАММНАЯ СИСТЕМА ДЛЯ РАСПОЗНАВАНИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

Алейкин В.В., Ладыженский Ю.В.
Донецкий национальный технический университет г.Донецк
Кафедра прикладной математики и информатики
E-mail: aleykin.vladislav@gmail.com

Аннотация

Алейкин В.В.*, Ладыженский Ю.В.** ***Программная распределенная система распознавания текстовой информации. *В статье рассматривается метод создания распределенной системы распознавания текстовой информации, классификация изображений и применение алгоритмов предобработки для удаления шума и коррекции изображений*

Общая постановка проблемы. В настоящий момент существует ряд программных продуктов, которые способны распознавать сканированные документы хорошего качества с текстовой информацией с вероятностью более 90%. Данные показатели достаточно велики, чтобы использовать такие программы, как в офисах, при распознавании документов, так и в промышленности, для контроля продукции и маркированных деталей.

При распознавании бумажных документов на практике ошибки в 10 буквах на одном листе не так значимы, как ошибки, получаемые в промышленной сфере. Различные условия окружающей среды, при получении снимков деталей, а так же возможные повреждения маркировки (сколы, царапины, пятна и др.) приводят к снижению вероятности корректного распознавания символа на детали. Существующие программные продукты распознавания текстовой информации способны убрать слабые помехи в виде зернистого шума, связанного с низким качеством съемки. Более крупные помехи существующие программные продукты не способны определить и убрать со снимков.

Постановка задач исследования. Важными для развития теории и практики распознавания являются Captcha-изображения. Captcha - это полностью автоматизированный публичный тест Тьюринга, чтобы отличать компьютеры от людей (рис. 1). По отношению к автоматическому распознаванию существуют понятия слабая CAPTCHA и прочная CAPTCHA. В числе слабостей, облегчающих распознавания: фиксированный шрифт, фиксированное положение символов, отсутствие искажений, отделение символов от фона с использованием цветового ключа или размытия по Гауссу, лёгкое отделение символов друг от друга и др.[1]

Рисунок 1 – Публичный тест Тьюринга - Captcha

Распознавание Captcha-изображений является сложной задачей. Существуют программные реализации, которые способны распознать только определенные Captcha-изображения, нет системы, которая может распознать абсолютно любую Captcha. Для распознавания используются методы восстановления изображения по содержимому на странице, подходы с использование нейронных сетей, шаблонное сопоставление, ручное распознавание и др.[2]

При разработке технологии распознавания текстовой информации ставится задача распознавания «Captcha-изображений»[3] с повышенным уровнем шума в виде

многочисленных отрезков кривых линий, и исследование эффективности применяемых методов при распознавании текстовой информации.

Структура программной системы включает модуль распознавания, который реализует алгоритм распознавания на основе многослойной нейронной сети, и модуль предобработки, реализующий серию алгоритмов для коррекции изображения и удаления посторонних шумов, негативно влияющих на процесс распознавания.

Решение задачи и результаты исследований. В модуле распознавания реализована нейронная сеть - многослойный персептрон[4]. Персептрон обладает двумя внутренними слоями, что повышает точность при распознавании.

Реализация нейронной сети включает два этапа:

1) Обучение нейронной сети на тестовых образах;
2) Распознавание реальных образов.

Обучение многослойного персептрона реализовано, как «обучение с учителем», и состоит в изменении весовых коэффициентов для каждой связи между слоями[4]. На вход подавались изображения отдельных символов с нанесенными слабыми помехами в виде дополнительных штрихов (рис. 2).

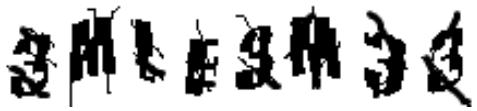

Рисунок 2 – Пример образа для обучения

Всего на вход предоставлено 8 различных наборов символов. В каждом наборе находится 26 символов английского алфавита и 10 цифр. Таким образом, нейронная сеть обучена на 288 символах. Каждый символ уникален – имеет набор помех, который не повторяется на других символах, шрифт, наклон и толщина символа варьируются случайным образом в одном из 8 наборов.

Обучение на таком наборе символов заняло примерно 45 минут, при этом частота процессора была 2.8Ghz и другие времяемкие процессы в это время не были запущены.

После обучения персептрон готов работать в режиме распознавания. В этом режиме персептрону предъявляются ранее неизвестные ему образы, и персептрон должен установить, к какому классу символов они принадлежат. При использовании обученного персептрона вероятность распознавания достигла 98% из 300 образов. Данный показатель является хорошим для такого небольшого обучающего множества. Но при использовании изображений с нанесенными помехами повышенной сложности (рис. 3а, 3б) нейронная сеть показала невысокую эффективность: вероятность распознавания составила 28% из 300 образов.

а б

Рисунок 3 – Пример образов с повышенным уровнем помех

Для решения проблемы, связанной с большим количеством помех на изображении разработан специальный модуль предобработки. Данный модуль выполняет преобразование цветного изображения в монохромное, удаление помех в виде кривых, сегментирование на отдельные символы.

На вход модулю подается цветное изображение, содержащее символы с наложенным шумом. Входное изображение подвергается бинаризации по яркости для отделения нужной информации от фонового цвета; подсчитывается количество значимых черных пикселей на

изображении и строится вертикальная гистограмма яркости (рис. 4); если количество значимых пикселей меньше установленного порогового значения (на гистограмме отсутствуют четко выделенные пики), то применяется алгоритм сегментирования изображения, использующий гистограмму яркости и высоту символа[5], иначе, если гистограмма содержит четко выделенные пики, то применяется алгоритм сегментирования по пикам вертикальной гистограммы яркости[6]. В данном алгоритме правильное пороговое значение очень важно, так как символы различаются по толщине написания, все символы разделяются на два класса. В первый класс входят все символы с жирным написанием, во второй – с обычным написанием. После описанных этапов на выходе получается монохромное изображение с удаленными помехами и шумом, на котором все символы отделены друг от друга.

На рис. 5 представлены результаты работы реализованного модуля (слева – входной образ, справа – обработанный модулем). Видно, что помехи достаточно хорошо удаляются с использованием описанного алгоритма.

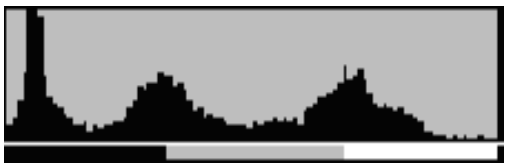

Рисунок 4 – Вертикальная гистограмма яркости

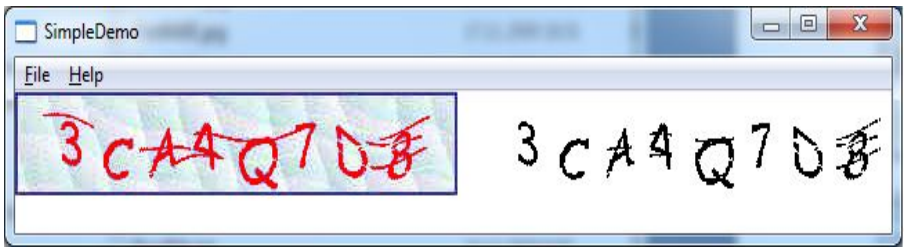

Рисунок 5 – Удаление помех

Для ускорения вычислений рассмотренные алгоритмы могут быть реализованы на распределенной вычислительной системе. Учитывая, что на вход подаются изображения двух различных классов, имеет смысл организовать распределенную систему так, чтобы часть процессоров занималась обработкой первого класса изображений, а другая часть – вторым классом[7]. На рис. 6 представлена структура распределенной системы. На вход подаются документы, требующие распознавания. Для каждого входного изображения определяется его класс и изображение передается определенной группе процессоров. После процесса сегментации обработанное изображение передается модулю распознавания, который на выходе предоставляет распознанную текстовую информацию для каждого изображения.

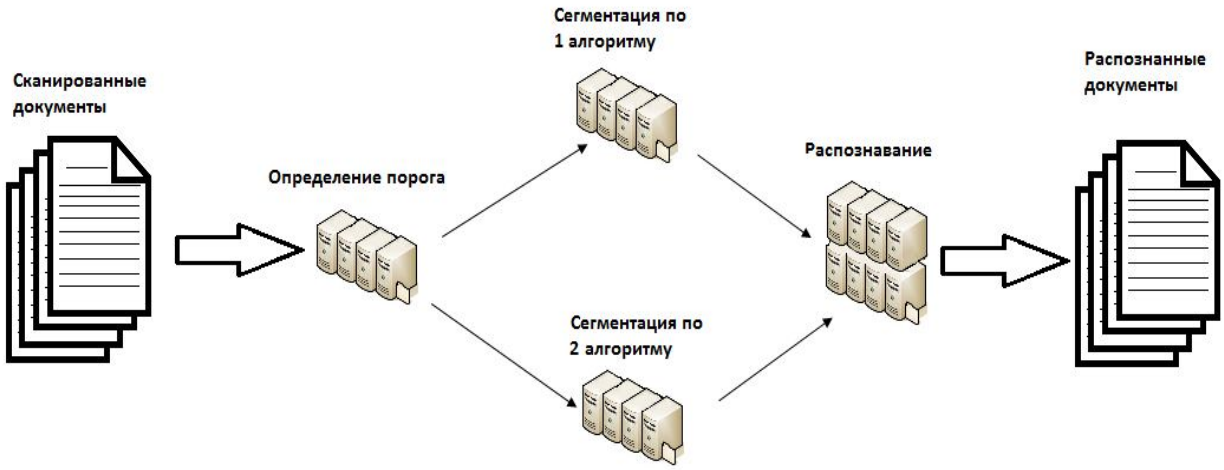

Рисунок 6 – Структура распределенной системы

Выводы. Разработаны алгоритмы обработки и распознавания текстовых изображений с помехами изображений. По разработанным алгоритмам реализован программный модуль предобработки изображений, снижающий уровень помех.

При тестировании разработанного алгоритма обработки и распознавания текстовой информации на 300 различных изображениях вероятность корректного распознавания составила 75%.

Для распознавания реализован многослойный персептрон в виде отдельного модуля. Разработана структура распределенной системы для обработки потока изображений. С использованием разработанной технологии распределения обработки образов предобработка выполняется на одной группе компьютеров, а распознавание на другой – это приводит к повышению быстродействия обработки изображений.

Список литературы

1. “CAPTCHA”. http://en.wikipedia.org/wiki/CAPTCHA (обращение 14.04.2010)
2. “Breaking da CAPTCHA theShockwaveRider” http://xain.hackerdom.ru/zine/online/issue0/Breaking%20Da%20CAPTCHA.html (обращение 14.04.2010)
3. «CAPTCHA Effectiveness» http://www.codinghorror.com/blog/2006/10/captcha-effectiveness.html (обращение 14.04.2010)
4. Шапиро Л., Стокман Дж. Компьютерное зрение; Пер с англ. – М.: БИНОМ. Лаборатория знаний, 2009. – 752с.
5. Форсат, Дэвид А., Понс, Жан. Компьютерное зрение. Современный подход. : Пер. с англ. – М. : Издательский дом «Вильямс», 2004. -928с.
6. Корнеев В.Д. Параллельное программирование в MPI. – 2-е изд., испр. – Новосибирск: Изд-во ИВМиМГ СО РАН, 2002. -215с.
7. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – М.: Горячая линия – Телеком, 2001. -382с.