Беззубов И.Д.
Научный руководитель – к.т.н. Е. Е. Канунова
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23

Реализация алгоритма восстановления распознанных символов, на изображениях старопечатных документов, полученных путем распознавания

Среди проблем формирования современной научной и образовательной среды гуманитарных наук одной из актуальных является создание информационных ресурсов на основе рукописных и старопечатных документов XII-XVII вв. Это подтверждается вниманием к различным сторонам ее решения специалистами в области гуманитарной информатики, историками, лингвистами, филологами, работниками музеев, архивов. Сделаны существенные шаги в области каталогизации, документирования, сохранения и визуализации этих ценных исторических источников на основе информационных технологий. В интернете растет число коллекций их электронных версий. В то же время, в большинстве случаев такие коллекции представляют собой цифровые изображения документов, что позволяет решать задачи их сохранения, визуализации, расширения доступа исследователей к ним, но ограничивает возможности содержательного информационного поиска и анализа с помощью современных компьютерных методов. Ограничения исследовательских возможностей связаны, прежде всего, с трудностями представления электронных версий рукописных и старопечатных книг в формате электронного текста, ввиду отсутствия эффективных систем распознавания.

Традиционным источником хранения информации оставался и остается «бумага» и остальные подобные материалы. Однако бумага имеет свойство портиться со временем, желтеет, пачкается, рвется, теряет свои свойства и, в конце концов, может просто рассыпаться. Соответственно информация, которая хранится на бумаге, частично или полностью исчезает. Возможности электронной техники расширяют возможности людей по обработке изображений.

Документы являются как вещественными, так и письменными доказательствами. Своеобразную группу составляют те из них, которые иногда совмещают в себе признаки вещественного и письменного доказательства - так называемые старые документы, т. е. документы, в которых произошли физико-химические изменения материалов письма.

Следовательно, задача распознавания и восстановления рукописного текста считается актуальной на сегодняшний день. Точное распознавание латинских символов в печатном тексте в настоящее время возможно только если доступны чёткие изображения, такие как сканированные печатные документы. Точность при такой постановке задачи превышает 99%, абсолютная точность может быть достигнута только путем последующего редактирования человеком. Проблемы распознавания рукописного «печатного» и восстановление, как отдельных символов, так и всего документа, а также печатных текстов других форматов (особенно с очень большим числом символов) в настоящее время являются предметом активных исследований.

Обычно траектория символа имеет разрывы, соответствующие отрыву пера от бумаги. Каждой траектории на изображении соответствует штрих – полоса черных точек шириной, равной диаметру пишущего инструмента. Давление пера на бумагу не постоянно, поэтому в разных точках траектории толщина штриха может быть различна. Штрих может иметь самопересечения, пересекаться с другими штрихами, накладываться на себя и на другие штрихи. За счет изменения площади соприкосновения пера с бумагой, а также за счет искажений, связанных с процедурой сканирования, штрих может быть существенно искажен на изображении – искажаются границы и возникают случайные разрывы штриха.

На сегодняшний день мы имеем две разные постановки вопросов восстановления дефектных символов, отличие которых связано с методом получения изображения. Изображение символа можно получить при сканировании документа, имеющего старопечатный текст. В данном случае входной информацией для задачи распознавания и последующего восстановления символов считаются изображения эталонных символов, совпадающие с распознанными буквами, и появляется

задача замены символа с дефектом. Альтернативный метод восстановления символа на изображении документа - это применение специализированных устройств, например, графический планшет. При этом изображение вносится в память компьютера в процессе написания символов; входной информацией для задачи считаются траектории перемещения пера, представляющие собой последовательности координат пера.

Большинство существующих методов решения задачи распознавания и восстановления символов включает следующие основные этапы: предобработка изображения документа, формирование массива распознанных символов или структурного представления, выбор символа, имеющий дефект, замена участка изображения с дефектным символом, на более качественный эталон. Набор признаков формируется по следующим видам данных, полученных на этапе предобработки: бинарная матрица, сглаженный граничный контур и скелет изображения. Такой подход позволил достичь высокой точности распознавания напечатанных и аккуратно написанных символов.

В данной работе предлагается подход к восстановлению как отдельных символов, так и документа полностью, основанный на распознавании образов символов с изображения документа и последующей замене символов, имеющих визуальный дефект.

Цель работы заключается в реализации алгоритма восстановления дефектных символов на изображениях старопечатных документов, полученных путем распознавания.

Основные задачи:

1. Провести предобработку изображений старопечатных документов;
2. Собрать базу изображений старопечатных эталонных символов;
3. Провести распознавание символов на отдельных участках документа, или полностью;
4. Определить дефектные символы и произвести восстановление, путем замены на эталонный аналог.