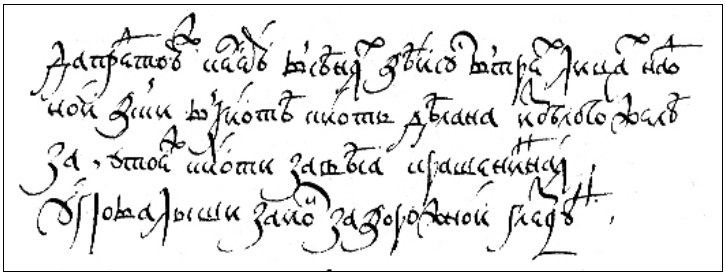
Зеленцов И.А.

Учебно-практические занятия по распознаванию древнерусской скорописи

1 Введение

В настоящее время важной задачей в вопросе сохранения культурного наследия страны является перевод памятников письменности в цифровую форму. Существует два способа цифрового представления письменного документа: цифровое растровое изображение документа, полученное путём его сканирования, а также электронный текст документа. Автоматизация процесса получения электронных документов второго вида решается использованием систем оптического распознавания (OCR). Однако, если речь идёт о таком классе документов, как древнерусская скоропись, применение современных систем распознавания оказывается невозможным. Это связано и с языком рукописей, претерпевшего к сегодняшнему дню значительные изменения, и со способом письма, принципиально отличного от типографской печати. Отсюда возникает необходимость разработки специализированной системы распознавания.



Целью предлагаемых учебно-практических занятий является введение обучающегося в задачу распознавания древнерусских скорописных документов XVI-XVII вв. Выполнение заданий позволяет на практике изучить особенности рассматриваемых текстов, получить представление о характере процесса их распознавания и возможных трудностях. Кроме того, в результате выполнения заданий формируется материал, пригодный для использования в качестве экспериментальной базы при разработке системы распознавания.

2 Состав рабочих материалов

Рабочие материалы организованны следующим образом:

- fonts
  - *.ttf
- images
  - 1.bmp
  - 2.bmp
  - ...
- texts
  - 1.rtf
  - 2.rtf
  - ...
- Методические материалы
  - Шаблон
    - letters
      - А
      - ...
      - Ять
    - words
  - Задание.doc

Каталог images содержит изображения сканированных страниц скорописных томов в формате BMP (RGB). Каждая страница представлена в отдельном файле, имеющем название <номер_страницы>.bmp.

Каталог texts содержит файлы в формате RTF, в которых представлены тексты, изображённые на соответствующих изображениях из каталога images.

В каталоге методические материалы представлены различные вспомогательные материалы. Подкаталог Шаблон содержит скелет иерархии каталогов, необходимых для выполнения заданий.

3 Порядок выполнения работы

3.1 Подготовка

Перед выполнением работы рекомендуется установить в операционной системе шрифты из каталога fonts. Тексты из каталога texts рекомендуется просматривать в начертании шрифта IzhitsaC.

Для каждой обрабатываемой страницы необходимо создать каталог, называемый соответствующим номером страницы. Далее этот каталог называется рабочим. В рабочий каталог копируется содержимое каталога Методические материалы\Шаблон. Кроме того, в корне рабочего каталога необходимо создать документ Adobe Photoshop с именем <номер_страницы>.psd. В этом документе в слое с названием Изображение должно располагаться обрабатываемое изображение из файла images\<номер_страницы>.bmp. Исходное изображение необходимо отфильтровать по порогу 196 и перевести в режим 1-битного кодирования цвета.

3.2 Задание 1

Из исходного изображения в слое Изображение необходимо выделить фрагменты, содержащие отдельные слова текста. Для идентификации слов рекомендуется пользоваться текстом, представленным в файле texts\<номер_страницы>.rtf. Следует обращать внимание на то обстоятельство, что некоторые буквы могут иметь надстрочное написание. В текстах каталога texts надстрочные буквы выделены курсивом. Знаки препинания следует оставлять без внимания. Однобуквенные слова обрабатываются наряду с прочими.

Выделение слова в слое Изображение должно производиться прямоугольной областью. Выделенный фрагмент должен включать в себя все элементы начертания слова. Допускается небольшое расстояние (2-5 пикселей) от крайних точек начертания слова до границ области выделения. Присутствие элементов других слов в области выделения допускается.

Каждый выделенный фрагмент-слово необходимо скопировать в отдельный слой, название которого должно представлять собой порядковый номер слова на странице. Размер и позиция слоя должны соответствовать размеру и позиции фрагмента в общем изображении. Для слова, имеющего перенос на новую строку, следует создать слой, состоящий из двух фрагментов изображения слова, причем фрагменты должны располагаться на тех же позициях, что и в исходном изображении (т.е. перенос фрагментов для «склейки» на данном этапе не требуется). Для слова, имеющего перенос на новую страницу, необходимо выделить лишь тот фрагмент, который располагается на данной странице.

3.3 Задание 2

В каждом слое, полученном при выполнении задания 1, необходимо выделить изображения отдельных букв. Для идентификации букв в слове также рекомендуется пользоваться текстовым представлением страницы.

Аналогично выделению слов, буквы должны выделяться прямоугольной областью. Выделение должно включать в себя все элементы начертания символа, при этом допускается присутствие элементов других букв.

Каждый выделенный фрагмент-буква должен быть скопирован в отдельный слой, вложенный в слой фрагмента-слова, и должен именоваться порядковым номером буквы в данном слове.

3.4 Задание 3

Каждый слой-слово необходимо скопировать во временное изображение. Во временном изображении необходимо выполнить следующие действия:

- Если слой состоит более чем из одного фрагмента (т.е. выделенное слово имело перенос), необходимо соединить фрагменты для получения целостного изображения слова;
- Удалить элементы других слов, попавшие в данный фрагмент. Удалению должны подвергаться только те элементы других слов, которые не соприкасаются с элементами данного слова;
- Увеличить размер изображения так, чтобы расстояние между крайними точками изображения слова и границами изображения составляло не менее 5 пикселей;
- Сохранить полученное изображение в файле
  <рабочий_каталог>\words\<имя_файла>.bmp,
  где
  <имя_файла> = <номер_страницы>-<номер_слова>,
  а <номер_слова> соответствует названию слоя, из которого получено данное изображение, т.е. является порядковым номером слова на странице.
  Например: после обработки шестого слова текста из исходного изображения images\34.bmp должен быть получен файл
  <рабочий_каталог>\words\34-6.bmp

Для каждого слоя-буквы необходимо выполнить аналогичные действия, скопировав слой в отдельное изображение:

- Удалить элементы других букв, попавшие в данный фрагмент. Удалению должны подвергаться только те элементы других букв, которые не соприкасаются с элементами данной буквы;
- Увеличить размер изображения так, чтобы расстояние между крайними точками изображения буквы и границами изображения составляло не менее 5 пикселей;
- Сохранить полученное изображение в файле
  <рабочий_каталог>\letters\<название_буквы>\<имя_файла>.bmp,
  где
  <название_буквы> соответствует букве на полученном изображении;
  <имя_файла> = <номер_страницы>-<номер_слова>-<номер_буквы>
  Например: На странице 15 (images\15.bmp) в пятом по порядку слове третьей буквой является буква К. Тогда в результате обработки данной буквы должен быть получен файл
  <рабочий_каталог>\letters\К\15-5-3.bmp

Примечание: сохраняемые изображения должны быть в режиме 1-битного кодирования цвета.

3.5 Выходные данные

В результате выполнения всех заданий для конкретной страницы (на примере обработки изображения 1.bmp) должен быть получен каталог с именем «1» следующего содержания :

- 1
  - letters
    - А
      - 1-1-1.bmp
      - 1-1-2.bmp
      - ...
      - 1-5-3.bmp
      - ...
    - Б
      - ...
    - ...
    - Ять
      - ...
  - words
    - 1-1.bmp
    - 1-2.bmp
    - ...
    - 1-5.bmp
    - ...
  - 1.psd

3.6 Дополнительные материалы

1. http://www.vgd.ru/STORY/skoropis.htm (в особенности главы 3, 4, 9);
2. Черепнин Л.В. Русская Палеография — М., 1956 г. - глава 4, §4