Автоматическая сегментация строк в изображениях рукописных документов

Захаров Егор Олегович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: eo.zakharov@gmail.com

Понятие строки является ключевым при работе с электронными архивами сканированных текстовых документов как печатных, так и рукописных. В данной работе рассматривается задача сегментации строк в изображениях рукописных документов, которая возникает при организации навигации по большим массивам изображений текста. Задача сегментации строк состоит в нарезке изображений текста на фрагменты, включающие ровно одну текстовую строку. Сложность этой задачи определяется тем, что в рукописных документах (черновиках, дневниках, записных книжках) мы не можем опираться на предположения о структуре строк, справедливые для печатных документов: например, об обязательном наличии междустрочных интервалов, о параллельности строк и единой их ориентации на странице. В случае рукописных документов эти предположения либо не выполняются, либо выполняются лишь частично. Данная работа выполняется в рамках исследования по автоматизации работы с электронными архивными документами русских писателей.

Предлагаемый подход к решению связан с построением геометрического скелета для изображения текста, который представлен, как объект одного цвета на фоне другого цвета (бинарным изображением). Скелетом (или срединной осью) такого изображения мы будем называть множество центров максимальных вписанных кругов изображения текста. Это множество представляет собой планарный граф. Задача сегментации строк рукописного текста сводится к выделению подграфов в полученном скелете, при этом подграфы должны быть: а) непересекающимися, б) каждый подграф должен соответствовать одной строке. Данную задачу можно рассматривать как задачу кластеризации.

Формально исходная задача разбивается на несколько подзадач:

$$\text{Image} \xrightarrow{preprocessing}$$

$$\text{Binary Image} \xrightarrow{skeletonization}$$

$$(V, E) \xrightarrow{clusterization} \{(V_l, E_l)\}_{l=1}^{L}$$

Где:

- Binary Image — двухцветное изображение
- (V, E) – планарный граф, V – множество его вершин, E – множество рёбер
- $V_l \in V, V_i \cap V_j = \varnothing, i \neq j$, $E_l \in E, E_i \cap E_j = \varnothing, i \neq j$
- L – число найденных кластеров в исходном графе (т.е. число строк в изображении)

Предлагаемый метод кластеризации рёбер включает в себя две стадии: разбиение графа на компоненты связности с последующим исключением из них рёбер, чья ориентация (средний угол наклона) отличается от доминирующего в компоненте, а также последующая кластеризация полученных компонент связности в строки по признаку близости этих компонент, а также среднему углу наклона.

Использование данного подхода предполагает свои особенности предшествующей фильтрации изображения. В реализованном методе помимо использования state-of-the-art техник бинаризации, учитывается специфика задачи: необходимость размытия больших связных участков текста в изображении, в связи с чем к нему применяется ряд линейных и нелинейных преобразований, обеспечивающих получение наименее зашумлённого графа скелета (V, E).

Совокупность предложенных методов является первым шагом на пути решения проблемы сегментации и поиска информации в рукописных документах, которая возникает при работе с электронными архивами текстов. Появление системы адаптивной строковой навигации позволит более эффективно обмениваться информацией и сэкономит время на её поиск.

Литература

1. Местецкий Л. М. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. М.: Физматлит, 2009.
2. Ntirogiannis K. et al. A combined approach for the binarization of handwritten document images // In Pattern Recognition Letters, 2012