Метод распознавания символов в изображениях тибетских манускриптов

Шолохова Татьяна Николаевна

Студент

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

E-mail: tanja200596@mail.ru

Потребность в преобразовании рукописных архивных документов в электронный вид возникла из-за огромного количества проблем, связанных с хранением материальных форм данных: отсутствие возможности быстрого доступа, сложность создания копий, неизбежные повреждения с течением времени.

Целью данной работы является исследование непрерывно-морфологического подхода к решению задачи оптического распознавания символов (optical character recognition, OCR), применительно к образцам древних тибетских рукописей.

Непрерывно-морфологический подход подразумевает, что правила идентификации основываются на геометрических свойствах, полученных путём аппроксимации растрового представления символов непрерывными геометрическими фигурами.

Выполненная работа включает: построение и обработку внутреннего скелета изображения текста, построение так называемого штрихового представления; классификацию и агрегирование штрихов скелета; выбор эвристических правил идентификации символа, основанных на интуитивных предположениях, автоматических правил идентификации, основанных на алгоритмах машинного обучения; выделение из скелетного представления символа информативных признаков для обучения алгоритмов машинного обучения; оценку качества полученных методов.

Первый шаг выполнения работы заключался в построении внутреннего скелета изображения текста. В геометрии и компьютерной графике под термином «скелет фигуры» понимают множество точек (пикселов, векселов), являющихся центрами максимальных вписанных кругов. Алгоритм построения скелета позволяет получить его в виде плоского геометрического графа [1].

Следующий шаг заключался в построении штрихов, на основе полученного скелета. Штрихи представляют собой модель следа пера. Агрегация — это некоторое объединение инцидентных рёбер скелета, при котором сохраняются основные свойства скелета. Агрегация нужна для того, чтобы построить признаковое описание, адекват-

но отражающее процесс письма. Был разработан двухэтапный метод агрегации. Первый этап основывается на жадной идее: пока в скелете есть пара инцидентных штрихов с близким к развёрнутому углом между ними, штрихи объединяются в один. Второй этап переборный: перебираются всевозможные цепочки графа, для каждой цепочки перебираются всевозможные сглаживания, после чего выбирается оптимальное сглаживание. В результате получается набор прямолинейных штрихов (Рис. 1).

Далее из изображения рукописи извлекались символы. Считалось, что один символ содержится в квадратной рамке, с центром в середине некоторого штриха. Для каждого символа производилось построение скелета, с дальнейшим агрегированием и выделялись признаки, на которых основывался дальнейший анализ.

Главным образом признаки опирались на типы штрихов, находящихся внутри символа-рамки. Штрихи делились на типы по местоположению (рамка равномерно разделялась на 9 секторов, штриху в соответствие ставился тот сектор, в который попадает его центр), по длине (короткие, средние, длинные) и по направлению (вертикальные, горизонтальные, диагональные, антидиагональные). Итого получалось $4 \times 9 \times 3 = 108$ возможных типов.

В качестве признаков использовались гистограммы распределения штрихов по типам, где все штрихи вносили одинаковый вес и аналогичные гистограммы, где каждый штрих вносил вес, пропорциональный своей длине. Дополнительно для всего подграфа-символа подсчитывались некотрые характеристики, такие как: длина минимального и максимального ребра, самый длинный путь, средняя длина всех рёбер, количество компонент связности и некоторые функции от данных величин.

Описанные выше признаки использовались для обучения алгоритмов классификации (логистическая регрессия, метод опорных векторов, случайные деревья). Работа классификаторов демонстрируется на распознавании символа «са» (Рис. 2.).

Выбранная метрика AUC_PR (площадь под precision recall кривой) учитывает несбалансированность данных. Наилучший результат получен алгоритмом случайных деревьев 0.835 (Рис. 3.). Проведённые исследования показываются, что предложенный подход является перспективным и его дальнейшее развитие может привести к разработке эффективных алгоритмов распознавания.

Иллюстрации

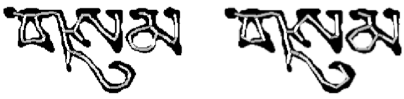

Рис. 1. Пример скелета до (слева) и после агрегирования (справа).

Рис. 2. Пример распознавания символа «са» на изображении древней рукописи.

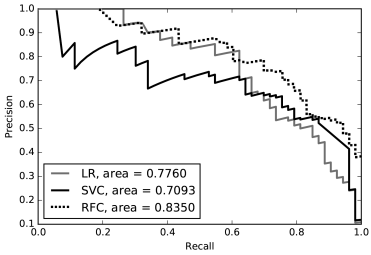

Рис. 3. Результаты работы классификаторов.

Литература

1. Местецкий Л. М. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры.
   Москва, Физматлит, 2009.
2. Местецкий Л. М. Скелет многоугольной фигуры — представление плоским прямолинейным графом.
   Санкт-Петербург, ГРАФИКОН-2010.
3. Xinbo Gao, Bing Xiao, Dacheng Tao. A survey of graph edit distance.
   FORMAL PATTERN ANALYSIS & APPLICATIONS, 2010.