РАЗРАБОТКА АВТОМАТИЧЕСКОЙ СИСТЕМЫ ДЛЯ ПРОВЕРКИ ПРАВИЛЬНОСТИ НАПИСАНИЯ СИМВОЛОВ

П.А. Хаустов
(г. Томск, Томский политехнический университет)
E-mail:exceibot@tpu.ru

IMPLEMENTATION OF AN AUTOMATIC SYSTEM FOR CHARACTER SPELLING VALIDATION

P.A. Khaustov
(Tomsk, Tomsk Polytechnic University)

Abstract. The algorithm for character topology composition has been proposed. Metrics for character graphs comparison has been suggested. The proposed algorithm has been implemented in character processing application and has been approved on MNIST handwriting characters database and writing characters examples from the forms of a unified state exam.

Keywords: optical character recognition, character topology, character spelling validation, automatic system, spelling assessment.

Введение. При подготовке школьников России к единому государственному экзамену, как правило, основной акцент делается на решение самих заданий. Однако организаторы единого государственного экзамена все чаще замечают недостаточную подготовленность школьников России в плане умения правильно заполнять бланки ЕГЭ, учитывая предложенные им примеры начертания символов. Школьники нередко пренебрегают правилами заполнения бланков и не следуют строго приведенным примерам.

Возникает закономерное предположение о целесообразности разработки автоматической системы для проверки качества начертания символов в бланках для выполнения экзаменационных заданий. К основным трудностям, которые возникают в процессе разработки подобной системы, можно отнести неоднозначность при выборе способа представления начертаний символов, аналогичную неоднозначность при выборе функции оценки схожести двух начертаний, а также необходимость разъяснения тестируемому, почему такое начертание имеет недостаточную степень схожести с предложенным ему примером начертания символа.

Предложенный метод. С учетом необходимости аргументирования оценки степени схожести с некоторым эталонным изображением, возникает необходимость в подходе, отличном от использования нечетких классификаторов, таких как искусственные нейронные сети, аппарат нечеткой логики или машины опорных векторов [1]. Для изображения символа замену нечетким классификаторам можно выполнить с использованием построения топологической модели графического представления символа. Одним из вариантов представления начертания символа является его представление в виде планарного графа, вершинами которого являются некоторые ключевые точки графического представления, а ребрами – соединяющие их участки графического представления. Ребра при таком представлении, как правило, не могут быть представлены в виде отрезков на плоскости. Каждое из таких ребер может быть представлено в виде некоторого количества последовательно соединенных отрезков, дуг и, возможно, эллиптических дуг.

Для получения информации о топологии начертания символа необходима предварительная скелетизация его графического представления [2]. Так как каждый из общеизвестных алгоритмов скелетизации, обладающих высоким быстродействием, имеет свои недостатки, было решено последовательно использовать два известных алгоритма: алгоритм утончения Зонга-Суня и алгоритм Ву-Цая. Первый из этих алгоритмов периодически допускает присутствие на итоговом изображении неутонченных элементов, второй – зачастую удаляет небольшие элементы графического представления символа. Вследствие чего было высказано предположение об использовании алгоритма Ву-Цая для устранения нежелательных необработанных участков, оставшихся после использования алгоритма Зонга-Суня.

Для получения топологической модели по уже утонченному изображению используется многократный запуск алгоритма Ли [3]. В роли вершин используются все пикселы, при-

надлежащие графическому представлению символа и не являющиеся фоном. С помощью набора эвристик обнаруживаются ключевые пикселы, которые последовательно удаляются перед очередным запуском алгоритма Ли, чтобы проанализировать расположение остальных пикселов относительно друг друга.

Полученную топологическую модель было решено сохранять в xml-файл, для визуализации такого формата было реализовано web-приложение с использованием javascript. Координаты всех вершин и узловых точек ломанных нормированы так, чтобы изображение целиком умещалось внутри квадрата с углами в точках (0; 0) и (1; 1).

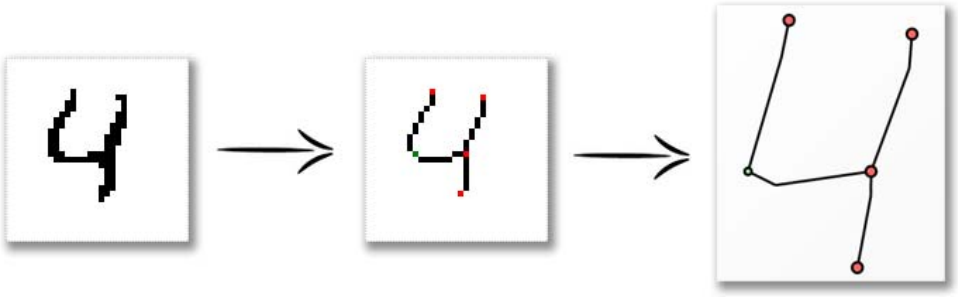

Рис. 1. Пример обработки графического представления символа

Для проверки некоторого начертания символа на соответствие указанному шаблону используется эвристический алгоритм нахождения паросочетания ребер двух топологических графов минимального веса. В качестве весов используется площадь области, заключенной между соответствующими ребрами топологической модели.

Полученные результаты. На данный момент производится апробация предложенного подхода на работах учащихся Томской области. В качестве эталонных изображений используются образцы начертаний из бланков ЕГЭ прошлых лет (пример таких изображений представлен на рис. 2).

Рис. 2. Образцы начертания символов из бланков ЕГЭ прошлых лет

При тестировании установлено, что линейная зависимость быстродействия алгоритма от количества пикселов на изображении подтверждается высоким быстродействием на практике (построение топологической модели занимает не более секунды на современных процессорах).

Предварительные результаты демонстрируют, что алгоритм выдает релевантные значения оценки степени схожести для обрабатываемых входных изображений.

Список литературы

1. Schantz, Herbert F., The history of OCR, optical character recognition – «Recognition Technologies Users Association», 1982. – 213
2. Роджерс Д., Алгоритмические основы машинной графики – М.: Мир, 1989. – С. 54-63.
3. Кристофидес Н. Теория графов. Алгоритмический подход – М.: Мир, 1978. – 145 с.