УДК 004.93'14

А. О. Сахарчук
Учреждение образования «Гродненский государственный университет имени Янки Купалы», Гродно

О ПРИМЕНЕНИИ АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА ПРИ РАСПОЗНАВАНИИ CAPTCHA

Введение. Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы (кластеры). Внутри каждого кластера в итоге должны оказаться «подобные» объекты, а объекты разных кластеров должны быть как можно более отличны друг от друга.

Применение кластерного анализа в общем виде сводится к следующим этапам: 1) отбор выборки объектов для кластеризации; 2) определение множества факторов и свойств, по которым будет дана оценка объектам в выборке. При необходимости нужно будет провести нормализацию значений этих факторов и свойств объектов; 3) вычисление коэффициента сходства между объектами; 4) применение метода кластерного анализа для создания кластеров.

CAPTCHA (англ. Completely Automated Public Turing test to tell Computers and Humans Apart) — компьютерный тест, используемый для того, чтобы определить, кем является пользователь системы: человеком или компьютером. В наиболее распространённом варианте CAPTCHA пользователь вводит символы, изображённые на рисунке (зачастую с добавлением помех), но так, чтобы машинное распознавание текста было максимально трудоёмким и невозможным.

Основная часть. Реализация приложений, связанных с автоматической обработкой данных из разных источников, находящихся на различных веб-ресурсах (веб-страницах, сайтах, порталах, страницах социальных сетей и т. д.) является очень актуальной и достаточной сложной задачей. Кроме самой сложности «парсинга» и обработки данных этих ресурсов существует и другая проблема. Эта сложность связана с нежеланием некоторых владельцев ресурсов давать доступ таким автоматизированным системам к своей информации. Причины могут быть различными: сильная нагрузка на сервер, нежелание делиться огромными объёмами информации и т. д. Для решения данной задачи была придумана система тестирования CAPTCHA, главной задачей которой является определение, является ли пользователь живым человеком или машиной. Но технологии в области распознавания изображений не стоят на месте. Стоит отметить, что в настоящее время некоторые «устаревшие» виды CAPTCHA можно распознавать, обучив этому компьютерное приложение.

Для исследований была выбрана CAPTCHA (рисунок 1), обладающая следующими сложностями при распознавании: 1) при каждой новой генерации цвета символов и цвета заднего фона могут отличаться. Цвета при этом имеют разницу в оттенках на различных участках изображения; 2) при каждой новой генерации символы могут иметь разный угол наклона и толщину; 3) в каждой генерации существует волнистая линия (главная помеха), которая пересекает все символы по горизонтали изображения и имеет такой же оттенок цвета, как и символы.

Рисунок 1 — Примеры исследуемой CAPTCHA

Также стоит отметить некоторые положительные факторы CAPTCHA (*см.* рисунок 1), которые важно применить для увеличения эффективности её распознавания. Большинство факторов были исследованы путём глубокого анализа достаточно большой выборки генераций CAPTCHA:

1) в каждой генерации присутствует ровно четыре символа латинского алфавита либо цифры. Иные символы либо их количество невозможны;

2) CAPTCHA содержит символы верхнего и нижнего регистра, однако ресурс, поддерживающий его, регистру значения не придаёт;

3) во избежание сложностей при распознавании человеком некоторые символы латинского языка и цифры отсутствуют (никогда не генерируются);

4) даже при наличии угла наклона и различий в толщине сам шрифт (очертания и шаблон) символов при каждой генерации остаётся неизменным;

5) ни в одной из генераций не было обнаружено никаких других помех, шумов, искривлений и графических эффектов, кроме горизонтальной волнистой линии;

6) линия (помеха) хоть и является кривой, но всё же имеет свою закономерность, которую можно описать математической формулой (для каждой генерации, имеющей свои коэффициенты, которые необходимо вычислить);

7) цвет всего фонового изображения (и его оттенков) сильно отличается от цвета (и оттенков) символов, притом количество пикселей фона всегда больше, чем количество пикселей символов;

8) в каждой генерации линия всегда касается левого и правого концов изображения, символы же напротив, никогда этого не делают.

Опираясь на все преимущества и недостатки (в смысле распознавания) выбранной CAPTCHA, можно выделить последовательность действий, необходимых для её распознавания:

1) для более удобного манипулирования данными было принято решение обрабатывать данные не самого изображения, а бинарной матрицы, в который за «0» принято считать пиксель заднего фона изображения, а за «1» — пиксель символов. Для создания такого рода матрицы был применён алгоритм кластеризации данных (фактором отбора является цвет пикселя), который распределяет символы на два кластера. Кластер с большим количеством пикселей — это фон изображения, а с меньшим — символы;

2) для более точного и эффективного распознавания данных необходимо отделить символы друг от друга и впоследствии работать с ними по отдельности. Для этого был применён алгоритм кластеризации данных, разбивающий матрицу на четыре кластера. За фактор сравнения были взяты координаты осей матрицы.

За помеху в данном случае можно считать волнистую линию, проходящую по горизонтали. Так как мы рассматриваем теперь каждую букву по отдельности, то для того, чтобы найти начало этой линии для следующего символа, необходимо передавать координаты от конца линии предыдущего символа. Началом линии для первого символа всегда будет самый левый пиксель изображения. (После вычисления линий на изображениях при помощи математических формул для сплайнов *n*-го порядка очень важно отметить, что необходим дополнительный алгоритм, который отслеживает пересечения линия с другими объектами. Это имеет критическую важность в том случае, если линия на некотором участке является одновременно и частью символа (рисунок 2).

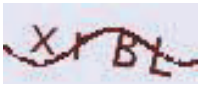

Рисунок 2 — CAPTCHA с линией, часть которой является частью символа

Для окончательного определения символа, который находится на изображении, необходимо методом последовательного сравнения сопоставлять полученное изображение с заготовленными заранее шаблонами символов и определять коэффициент попадания. Символ, имеющий наиболее высокий коэффициент попадания, считается тем, который находится на изображении.

Заключение. За время исследования были поставлены задачи изучения и нахождения решения для распознавания CAPTCHA определённого вида. Был предложен вариант использования алгоритмов кластерного анализа для фильтрации цвета и разделения символов изображения на отдельные элементы.

На данном этапе распознавание CAPTCHA реализовано с точностью в 25% и нуждается в некоторой доработке. Главными рисками падения эффективности являются алгоритмы разделения символов и удаления линии.

УДК 004.922

М. Ю. Сеч, А. И. Калько
Учреждение образования «Барановичский государственный университет», Барановичи

НАВИГАЦИОННАЯ СИСТЕМА ПОМЕЩЕНИЯ С ПРИМЕНЕНИЕМ QR-КОДОВ

Введение. В последнее время всё более актуальной становится проблема навигации внутри помещений, а также предоставления посетителям услуг, основанных на их местоположении (LBS, Location-based service) и предпочтениях. Здания становятся всё более объёмными и нередко имеют довольно сложную структуру, ориентироваться в которой могут лишь те, кто постоянно посещает такие здания, а для неподготовленного человека ориентирование в таких местах превращается в пытку.

Кроме того, решения, применяемые в indoor-навигации (навигации внутри помещений), помогают и в ориентировании вне зданий, на улице — там, где в условиях плотной застройки использование систем спутниковой навигации затруднено (нет спутников в прямой видимости, присутствует только отражённый/ослабленный/зашумленный сигнал GPS/Глонасс и т. д.). Особенно эта проблема актуальна для Японии с высокой плотностью городской застройки.

Основная часть. Основной недостаток систем спутникового позиционирования — проблематичность их применения в закрытых помещениях, в результате чего приходится искать иные пути решения проблемы indoor-навигации. Их несколько [1]:

1) навигация по вай-фай. Используется уже существующая инфраструктура сетей связи — точки беспроводных сетей вай-фай, и это наименее затратный вариант. Методика определения координат следующая: устройство пользователя сканирует доступные точки доступа вай-фай, затем информацию о них отправляет на сервер, где эти данные по базе данных сопоставляются с координатами этих точек доступа, по которым и вычисляются координаты пользователя. К сожалению, координаты точек вай-фай точно не известны, плюс могут меняться (перенесли вай-фай точку в другое место или заменили её на другую — координаты уже оказываются неверными). Точность при таком подходе оставляет желать лучшего (погрешность — до 25 м! А при использо-