УДК 004.932

АНАЛИЗ ТЕХНОЛОГИИ OCR РАСПОЗНАВАНИЯ ТЕКСТА НА ИЗОБРАЖЕНИЯХ

студент гр.814301 Шакун В.А.
Научный руководитель – к.т.н., доцент Ролич О. Ч.
Белорусский государственный университет информатики и радиоэлектроники
Минск, Беларусь

Вычислительные системы существуют не один десяток лет, и главной целью их создания являлась возможность замены человека и выполнения за него трудоемкой работы. К подобным системам относятся системы распознавания текстовой информации [1].

Алгоритмы распознавания текста используется в различных сферах. Они требуются для оцифровки старых книг, перевода текста изображения в электронный вид, облегчают процессы распознавания почтового индекса и идентификационного номера паспорта.

Распознавание текста – задача сложная для реализации. Человек для этого задействует комплекс знаний и опыта, выделяя текст из совокупности сигналов органов чувств, затем характерные признаки символов, и на основании собственного опыта делает вывод о значении символа и всего текста.

Разработчики программ распознавания текста сталкиваются с проблемами наложения символов друг на друга, их похожести в различных языках, низкого качества изображения, а также наличия шума на изображении. До сих пор не разработано программы, которая обеспечивала бы полную достоверность распознавания, поэтому в процесс распознавания символов и по сей день требуется вмешательство человека.

Целью данной работы является исследование технологии OCR (optical character recognition) для распознавания текстовой информации на изображениях. Объектом исследования является программная библиотека tesseract, основанная на технологии OCR.

Системы OCR состоят из следующих основных блоков, предполагающих аппаратную или программную реализацию и определяющих последовательность шагов обработки и анализа изображений: блок сегментации (локализации и выделения) элементов текста; блок предобработки изображения; блок выделения признаков; блок распознавания символов; блок постобработки результатов распознавания [2, 3].

Сначала осуществляется выделение текстовых областей, строк и разбиение связных текстовых строк на отдельные знакоместа, каждое из которых соответствует одному текстовому символу.

После разбиения (а иногда до или в процессе разбиения) символы, представленные в виде двухмерных матриц пикселей, подвергаются сглаживанию, фильтрации с целью устранения шумов, нормализации размера, а также другим преобразованиям с целью выделения образующих элементов или численных признаков, используемых впоследствии для их распознавания.

Распознавание символов происходит в процессе сравнения выделенных характерных признаков с эталонными наборами и структурами признаков, формируемыми и запоминаемыми в процессе обучения системы на эталонных и/или реальных примерах текстовых символов.

На завершающем этапе смысловая или контекстная информация может быть использована как для разрешения неопределенностей, возникающих при распознавании отдельных символов, обладающих идентичными размерами, так и для корректировки ошибочно считанных слов и фраз в целом.

При поступлении изображения на распознавание могут быть различные начальные условия: шумы, неправильная расположенность изображения, смазанность и

другое. Именно поэтому предобработка является важным этапом в процессе распознавания символов и позволяет производить сглаживание, нормализацию, сегментацию и аппроксимацию отрезков линий. Под сглаживанием в данном случае понимается большая группа процедур обработки изображений. В частности, широко используются морфологические операторы заполнения и утончения. С помощью заполнения можно устранить небольшие разрывы и пробелы. Утончение представляет собой процесс уменьшения толщины линии, в которой на каждом шаге области размером в несколько пикселей ставится в соответствие только один пиксель «утонченной линии».

В технологии OCR используется и алгоритм бинарной фильтрации как «стирание бахромы». Его суть заключается в последовательном стирании крайних элементов (например, крайнего верхнего, нижнего, левого и правого пикселей) на неровностях у границ символа, которые чаще всего мешают точному их определению, размеров символа и дальнейшему его распознаванию по контурному признаку.

Если в качестве апертуры фильтра выбрана окрестность второго порядка размером 3×3, то под крайним верхним пикселем понимается такой пиксель, в апертуре 3×3 которого наблюдаются следующие сочетания:

$$\begin{pmatrix}0&0&0\\0&1&0\\1&0&0\end{pmatrix},\begin{pmatrix}0&0&0\\0&1&0\\0&1&0\end{pmatrix},\begin{pmatrix}0&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}0&0&0\\0&1&0\\1&1&0\end{pmatrix},\begin{pmatrix}0&0&0\\0&1&0\\0&1&1\end{pmatrix},\begin{pmatrix}0&0&0\\0&1&0\\1&1&1\end{pmatrix},$$

где 0 обозначает элемент фона, 1 – элемент символа. С помощью поворота перечисленных матриц на 90°, 180° и 270° получаются определения соответственно крайнего левого, крайнего нижнего и крайнего правого пикселей.

При фильтрации с целью стирания бахромы все определенные таким образом краевые пиксели стираются. Стираются также изолированные пиксели, не имеющие соседних пикселей символа в апертуре 3×3. Остальные пиксели, не являющиеся крайними или изолированными, переносятся на отфильтрованное изображение без изменения.

По аналогии с описанным стиранием бахромы из единиц можно ввести стирание бахромы из нулей. При этом «краевые нули», апертуры которых соответствуют представленным выше матрицам с заменой единиц на нули и нулей на единицы, также «стираются», то есть замещаются единицами.

Алгоритм комбинированного стирания бахромы одновременно оперирует единицами по символу и нулями по фону. После использования данного алгоритма значительно увеличивается вероятность точного распознавания символов.

Для неправильно ориентированных изображений символов применяется геометрическая нормализация, отвечающая за устранение наклонов, перекосов отдельных символов и за корректирование их длины и ширины.

Процедура сегментации осуществляет разбиение изображения текста на отдельные объекты-сегменты: строки, слова и образы-символы. Данное решение эффективно лишь в случаях, когда символы текста не перекрывают друг друга. Слияние символов может быть вызвано типом шрифта, которым набран текст, плохим разрешением печатающего устройства или высоким уровнем яркости, выбранным для восстановления разорванных символов.

Под аппроксимацией отрезков линий понимается составление графа описания символа в виде набора вершин и прямых ребер, которые непосредственно аппроксимируют цепочки пикселей исходного изображения. Данная аппроксимация осуществляется для уменьшения объема данных и может использоваться при распознавании, основанном на выделении признаков, описывающих геометрию и топологию изображения.

Одной из самых сложных задач в распознавании образов является выделение признаков для каждого из них. На данный момент существует большое количество систем признаков. Проблема заключается в выделении именно тех признаков, которые позволят эффективно отличать один класс символов от всех остальных в данной конкретной задаче.

Так, метод сравнения образа с эталоном считается достаточно эффективным в распознавании символов. В этом случае определяется степень сходства между образом и каждым из эталонов. Классификация тестируемого изображения символа происходит по методу ближайшего соседа. Тем не менее, у данного корреляционного метода есть недостаток: любые помехи на изображении могут помешать грамотному распознаванию символа, поэтому применяются другие, специальные способы сравнения образов.

В статистической группе методов выделение признаков осуществляется на основе анализа статистических распределений точек. Известные методики этой группы используют вычисление моментов и подсчет пересечений. Моменты различных порядков с успехом используются в самых различных областях машинного зрения в качестве дескрипторов формы выделенных областей и объектов. В случае распознавания текстовых символов, в качестве набора признаков применяются значения моментов совокупности «черных» пикселей относительно некоторого выбранного центра. Наиболее общеупотребительными в приложениях такого рода являются построчные, центральные и нормированные моменты. Для цифрового изображения, хранящегося в двумерном массиве $f[][]$, построчные моменты являются функциями координат каждой точки изображения следующего вида:

$$m_{p,q} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} x^p y^q f[x][y],$$

где $p, q = 0, 1, ..., \infty$, M и N являются размерами изображения по вертикали и горизонтали, $f[x][y]$ – яркость пикселя в точке (x,y) на изображении.

Центральные моменты являются функцией расстояния точки от центра «тяжести» символа:

$$m_{p,q} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} (x-\bar{x})^p (y-\bar{y})^q f[x][y],$$

где $(\bar{x}, \bar{y})$ – координаты центра «тяжести» сегментированной области.

Нормированные центральные моменты получаются в результате деления центральных моментов на моменты нулевого порядка.

Следует отметить, что строковые моменты, как правило, обеспечивают более низкий уровень распознавания. Центральные и нормированные моменты предпочтительны вследствие их большей инвариантности к преобразованиям изображений.

В методе пересечений признаки формируются путем подсчета того, сколько раз и каким образом произошло пересечение изображения символа с выбранными прямыми, проводимыми под определенными углами. Этот метод часто используется в коммерческих системах благодаря его инвариантности к дисторсии и небольшим стилистическим вариациям написания символов, а также обладает достаточно высокой скоростью и не требует высоких вычислительных затрат.

Метод зон предполагает разделение площади рамки, в которую заключен символ, на области и последующее использование плотностей точек в различных областях в качестве набора характерных признаков.

В методе матриц смежности в качестве признаков рассматриваются частоты совместной встречаемости «черных» и «белых» элементов в различных геометрических комбинациях.

Метод характеристических мест использует в качестве признака число раз, которое вертикальный и горизонтальный векторы пересекают отрезки линий для каждой светлой точки в области фона символа.

К методам распознавания текста на базе интегральных преобразований относятся методы на основе интегралов Фурье и Фурье-преобразования, часто используемого как при распознавании символов, так и в работе со звуковыми файлами. Среди современных технологий распознавания, основанных на преобразованиях, выделяются методы, использующие Фурье-дескрипторы символов, а также частотные дескрипторы границ. Преимущества методов, использующих преобразование Фурье-Меллина, связаны с тем, что они обладают инвариантностью к масштабированию, вращению и сдвигу символа. Основной недостаток этих методов заключается в нечувствительности к резким скачкам яркости на границах. Например, по спектру пространственных частот сложно отличить символ «O» от символа «Q». В то же время при фильтрации шума на границах символа это свойство может оказаться полезным.

В методах распознавания текста на основе структурных составляющих структурные признаки обычно используются для выделения общей структуры образа. Они описывают геометрические и топологические свойства символа. Проще всего представить идею структурного распознавания символа текста применительно к задаче автоматического считывания почтовых индексов. В этом случае расположение каждого возможного отрезка-штриха уже заранее известно и различия символов заключаются в наличии или отсутствии целого штриха.

При анализе структурных составляющих сложных текстов дополнительную информацию вносят образы штрихов. Они помогают при определении характерных особенностей изображения: концевых точек, точек пересечения отрезков, замкнутых циклов. Так, если матрица, содержащая утонченный символ, разделена на девять прямоугольных областей в виде сетки 3×3, каждой из которых присвоен буквенный код от «A» до «I» (или от 0 до 8), то символ может рассматриваться как набор штрихов. При этом штрих, соединяющий некоторые две точки в начертании символа, может являться линией (L) или кривой (C). Штрих считается отрезком (дугой) кривой, если его точки удовлетворяют выражению [4]

$$ABS\left|\frac{\sum_{i=1}^{n} ax_i + by_i + \frac{c}{\sqrt{a^2+b^2}}}{n}\right| > 0.69.$$

В противном случае считается, что это прямолинейный отрезок. В данной формуле (x_i, y_i) является точкой, принадлежащей штриху; $ax + by + c = 0$ – уравнение прямой, проходящей через концы штриха; коэффициент 0.69 получен опытным путем. Далее символ может быть описан набором своих отрезков и дуг. Например, запись {«ALC», «ACD»} означает наличие прямой, проходящей из области «A» в область «C», и кривой, проходящей из области «A» в область «D» соответственно.

Основное достоинство структурных методов распознавания определяется их устойчивостью к сдвигу, масштабированию и повороту символа на небольшой угол, а также – к возможным дисторсиям и различным стилевым вариациям и небольшим искажениям шрифтов.

В существующих OCR-системах используются разнообразные алгоритмы классификации, то есть отнесения признаков к различным классам объекта. Они

существенно различаются в зависимости от принятых наборов признаков и применяемой по отношению к ним стратегии классификации.

Для классификации по признаку формируется набор векторов признаков, выступающих в роли эталонных, для каждого символа. На стадии обучения программы разработчик загружает большое количество образцов начертания символов, для чего, в том числе, используются словари. При этом для каждого образца система выделяет признаки и сохраняет их в виде соответствующего вектора признаков. Таким образом определяется кластер или класс в виде набора векторов признаков, описывающих символ.

В процессе использования системы OCR может появиться необходимость дополнительного обучения и задания новых векторов признаков или увеличения базы существующих кластеров.

На этапе распознавания в случае анализа по кластерам необходимо определить вектор признаков данного символа, а затем сравнивать с каждым классом, полученным ранее. Совпадение векторов признаков означает, что символ, наиболее вероятно, будет верно распознан. Алгоритмы классификации основаны на определении степени близости набора признаков рассматриваемого символа к каждому из классов. Правдоподобие получаемого результата зависит от выбранной метрики пространства признаков. Наиболее известной метрикой признакового пространства является традиционное евклидово расстояние:

$$D_j^E = \sqrt{\sum_{i=1}^{N} \left(F_{j,i}^L - F_i^L\right)^2},$$

где $F_{j,i}^L$ – i-й признак из j-го эталонного вектора; F_i^L – i-й признак тестируемого изображения символа.

При классификации по методу ближайшего соседа символ будет отнесен к классу, вектор признаков которого наиболее близок к вектору признаков тестируемого символа. Недостатком данного метода является то, что при увеличении набора классов и признаков значительно снижается скорость распознавания символа.

Одна из методик, позволяющих улучшить метрику сходства, основана на статистическом анализе эталонного набора признаков. При этом в процессе классификации более надежным признакам отдается больший приоритет:

$$D_j^E = \sqrt{\sum_{i=1}^{N} w_i \left(F_{j,i}^L - F_i^L\right)^2},$$

где w_i – вес i-го признака.

Другая методика классификации, требующая знания априорной информации о вероятностной модели текста, основана на использовании формулы Байеса. Из правила Байеса следует, что рассматриваемый вектор признаков принадлежит классу «j», если отношение правдоподобия λ больше, чем отношение априорной вероятности класса j к априорной вероятности класса i.

В достоверных системах OCR процесс распознавания не является достаточным и конечным, так как для верного растолкования целого набора символов и слов необходимо понимать контекст.

Существует множество OCR-приложений, использующих глобальные и локальные позиционные диаграммы, триграммы, n-граммы, словари и различные сочетания всех этих методов. Наибольший интерес представляют два подхода к решению данной задачи: словарь и набор бинарных матриц, аппроксимирующих структуру словаря.

Доказано, что словарные методы являются одними из наиболее эффективных при определении и исправлении ошибок классификации отдельных символов [4, 5]. При этом после распознавания всех символов некоторого слова словарь просматривается в поисках слова с учетом того, что оно, возможно, содержит ошибку. Если слово найдено в словаре, это не говорит об отсутствии ошибок. Ошибка может превратить одно слово, находящееся в словаре, в другое, также входящее в словарь. Такая ошибка не может быть обнаружена без использования смысловой контекстной информации: только она может подтвердить правильность написания. Если слово в словаре отсутствует, считается, что в слове допущена ошибка распознавания. Для исправления ошибки прибегают к замене такого слова на наиболее похожее слово из словаря. Если в словаре найдено несколько подходящих кандидатур для замены, то исправление не производится. В случае такого проблемного распознавания некоторые системы имеют реализованную возможность выбора подходящего к контексту слова пользователем.

Недостатком данного метода является то, что из-за объемности словаря, процесс поиска требуемого слова занимает достаточно много времени. Из-за этого весь процесс постобработки становится неэффективным.

Некоторые разработчики с целью преодоления трудностей, связанных с использованием словаря, пытаются выделять информацию о структуре слова из самого слова. Такая информация говорит о степени правдоподобия *n*-граммов (символьных последовательностей, например, пар или троек букв) в тексте. *n*-граммы также могут быть глобально позиционированными, локально позиционированными или вообще непозиционированными. Например, степень достоверности непозиционированной пары букв может быть представлена в виде бинарной матрицы, элемент которой равен 1, тогда и только тогда, когда соответствующая пара букв имеется в некотором слове, входящем в словарь. Позиционная бинарная диаграмма $D_{i,j}$ является бинарной матрицей, определяющей, какая из пар букв имеет ненулевую вероятность возникновения в позиции (*i, j*). Набор всех позиционных диаграмм включает бинарные матрицы для каждой пары положений.

Таким образом, технология OCR позволяет достаточно точно распознавать текст на изображениях, кроме вариантов, когда непосредственно текст искажен. Это могут быть случаи перекрытия символами друг друга или схожести образов одних символов на другие (например, сочетание букв «LI» может быть распознано, как «U»). Но до сих пор не разработана технология, позволяющая в автоматизированном режиме распознать текст без погрешностей, поэтому в данном вопросе ещё требуется вмешательство человека.

Литература

1. Бройдо, В. Л. Вычислительные системы, сети и телекоммуникации / В. Л. Бройдо. — СПб.: Питер, 2004. – 703 с.
2. Визильтер, Ю. В. Обработка и анализ цифровых изображений с примерами на LabVIEW IMAQ Vision / Ю. В. Визильтер, С. Ю. Желтов, В. А. Князь и др. – М.: ДМК-Пресс, 2009. – 465 с.
3. Волкова, М. А. Методы обработки и распознавания изображений / М. А. Волкова, В. Р. Луцив. – СПб: Университет ИТМО, 2016. – 40 с.
4. Липкина, А. Распознавание текста по структуре скелета букв / А. Липкина. – М.: МГУ им. Ломоносова, 2018. – 31 с.
5. Суясов, Д. И. Разработка алгоритмов распознавания текста на основе клеточных автоматов / Д. И. Суясов. – СПб.: ИТМО, 2007. – 88 с.