УДК 004.352.243

СНИЖЕНИЕ КОЛИЧЕСТВА ОШИБОК РАСПОЗНАВАНИЯ СКАНИРОВАННЫХ РУКОПИСНЫХ ТЕКСТОВ

***Львович И.Я.**, д.т.н., заместитель декана факультета информатики Панъевропейского университета, e-mail: komkovvivt@yandex.ru;*
***Львович Я.Е.**, д.т.н., президент Воронежского института высоких технологий, e-mail: komkovvivt@yandex.ru;*
***Мозговой А.А.**, аспирант Воронежского института высоких технологий, e-mail: komkovvivt@yandex.ru;*
Преображенский А.П., д.т.н., профессор Воронежского института высоких технологий, e-mail: app@vivt.ru;
***Чопоров О.Н.**, д.т.н., проректор по научной основе Воронежского института высоких технологий, e-mail: komkovvivt@yandex.ru.*

THE DECREASE IN THE NUMBER OF RECOGNITION ERRORS OF SCANNED HANDWRITTEN TEXTS

Lvovich I., Lvovich Ya., Mozgovoi A., Preobragenskiy A., Choporov O.

One of the most popular approaches for handwriting recognition is the representation of entire words in sequences of symbols of the Markov chain. The set extracted from the images of the symbols is analyzed for compliance with a pre-prepared word patterns (model templates). The word whose model has the highest probability of formation of the analyzed sequences recognized the target. The variability of cursive writing words leads to the need of the analysis extracted from the image sequence of characters with models generated for words consisting of different numbers of digits. In the case where the analyzed word different from the word used for the model-only template, model template longer words earns you a mathematical advantage over the model of the shorter words, leading to recognition errors. The paper proposes a solution to the problem based on normalization of the horizontal dimensions of the images of handwritten words.

Key words: OCR, optical recognition, handwriting, HMM.

Ключевые слова: оптическое распознавание, рукописный текст, оконное сканирование, СММ.

Одним из популярных подходов, применяемых для распознавания рукописного текста, является представление изображений целых слов в виде последовательностей символов марковской цепи. Набор извлекаемых из изображений символов анализируется на предмет соответствия заранее подготовленным моделям слов (модели-шаблоны). Слово, модель которого обладает наибольшей вероятностью формирования анализируемой последовательности, признаётся искомым. Вариативность написания рукописных слов приводит к необходимости анализа извлекаемой из изображения последовательности символов моделями, сформированными для слов, состоящих из разного количества символов. В случае, когда анализируемое слово отличается от слова, используемого для модели-шаблона только окончанием, модель-шаблон более длинного слова получает математическое преимущество над моделью более короткого слова, что приводит к ошибкам распознавания. В статье предлагается решение описанной проблемы на основе нормирования горизонтальных размеров изображений рукописных слов.

Введение

Распознавание текста на изображениях – очень актуальная тема для исследований, которые позволяют решать ряд научных и прикладных задач. Современные методы распознавания символов используются для решения широкого круга задач, как офисных, так и специализированных, например, распознавание изображений маркировки оборудования и др. Существует много методов для распознавания текста: метод структурных фреймов, метод биоалгоритмов анализа изображений, метод геометрических моментов, метод дескрипторов Фурье, метод вейвлет-преобразования, метод главных компонент, метод шаблонов и др. Однако, вопросы, которые связаны с распознаванием рукописного текста, особенно для систем с большой нагрузкой, исследованы не до конца.

Условно разделив текстовую информацию на печатную и рукописную и проанализировав достижения в области её оптического распознавания, несложно заметить, что качество распознавания печатных символов значительно лучше качества распознавания символов рукописных. Хотя источником информации в обоих случаях является человек, очевидно, что машина лучше оперирует данными, изначально введёнными с клавиатуры. На первый взгляд, это не кажется таким уж удивительным, но если вспомнить, что первые попытки научить машину (далее компьютер) мыслить, в том числе делать индуктивные выводы при анализе образов, были предприняты ещё в пятидесятых годах прошлого века, то разница в результатах становиться не такой очевидной. А ведь работа с первой моделью нейросети – перцептроном [2] – была начата Френком Розенблаттом примерно в то же время, что и создание Джоном Маккарти символьного языка программирования Лисп [3]. Оказалось, что значительно проще написать программу,

пользуясь ограниченным набором логически связанных операторов, и «научить» компьютер, например, играть в шахматы лучше ведущих гроссмейстеров, чем «научить» тот же компьютер мыслить хотя бы на уровне, позволяющем однозначно трактовать сочетания нескольких штрихов рукописного текста.

Первый патент на метод оптического распознавания был выдан более восьмидесяти лет назад. С тех пор качество распознавания печатных текстов постоянно улучшается и в данный момент находится на вполне приемлемом уровне. Иначе обстоит дело с распознаванием рукописного текста. Прямое копирование методик, применяемых для печатных символов, не даёт значимого результата, а попытки альтернативных подходов наталкиваются на многочисленные препятствия. Хорошим результатом на сегодня считается преодоление пятипроцентного порога в количестве ошибочно идентифицированных символов. Это равнозначно появлению в каждой строке текста двух-трёх ошибок, что недопустимо по причине больших трудозатрат на их последующее выявление и исправление.

Помимо того, что весь текст можно поделить на рукописный и печатный, рукописный текст, в свою очередь, также делится на две большие категории по методу извлечения из него информации для последующего распознавания. Это распознавание текста в процессе его написания «на лету» и распознавание текста, извлечённого из изображения. И в том и другом случаях могут применяться одинаковые методики, хотя и с разным успехом.

Если провести анализ способов написания рукописного и печатного текстов, то можно убедиться, что они сильно отличаются.

Каждое слово рукописного текста представляет собой комбинацию линий, которые получаются при движении пишущего узла от момента начала его движения по бумаге до момента завершения. Информация, полученная в процессе этого движения: траектория движения, скорость, сила нажатия на пишущий узел (при наличии такой возможности) и т.д. – используется для распознавания написанного текста. Отрезки линий внутри траектории движения могут использоваться вместе с информацией об их длине, либо рассматриваться как отрезки фиксированного размера. Это зависит от подхода, который будет использоваться в дальнейшем для распознавания написанного. Чаще других применяются методы опорных векторов (SVM-based approach) и скрытые марковские модели (СММ). Метод опорных векторов применялся, например, для группы романских [4], арабских [5], тайских [6] языков и арабских цифр [7]. СММ использовались при распознавании тайского [6], английского [8], арабского [9] и многих других языков.

Информация о характере движения пишущего узла позволяет значительно улучшить результаты распознавания, что происходит не только потому, что мы получаем информацию, которая является дополнительной к оптической информации. Улучшению способствует также факт неоднократного прохождения пишущего узла по одной и той же траектории.

Online распознавание слитного рукописного текста доступно в Windows 7, начиная с версии Home Premium [10]. Также существует ряд узкоспециализированных коммерческих продуктов, предоставляющих такую возможность (например, PenReader [11] компании Paragon Software). Качество получаемого результата позволяет использовать данное программное обеспечение для широкого круга задач.

Online распознавание крайне удачно вписывается в работу современных гаджетов. Мобильные устройства, такие как КПК, коммуникаторы, смартфоны должны удовлетворять двум взаимоисключающим требованиям. С одной стороны, иметь минимальные габаритные размеры, определяемые их «мобильностью», а с другой – размер дисплея этих устройств должен быть как можно больше, что необходимо для удобства их использования. Оставив от всего устройства только сенсорный дисплей, эту типичную для ТРИЗ задачу призводители успешно решили [12], правда, при этом пришлось пожертвовать удобством ввода текстовой информации. Текстовую информацию предлагается вводить посредством нажатий на виртуальные кнопки нарисованной клавиатуры.

Обобщив применяемые методы, можно заметить, что в основе каждого из них лежит попытка однозначной идентификации отдельных символов, а не морфем, которые используются только на этапе поиска возможных орфорграфических ошибок. Если для идентификации печатного символа этот подход обеспечивает вполне приемлемый результат, то в случае с рукописным вводом, когда разделение слова на отдельные символы представляет собой нетривиальную задачу, данный метод, очевидно, ошибочен.

«Оконное» сканирование

Использование для распознавания целых рукописных слов марковского моделирования [13], а именно: лево-правой модели Бакиса [14] – является на сегодня одним из самых популярных подходов. Марковское моделирование предполагает извлечение из изображений рукописных слов последовательностей символов разнообразными способами.

Наиболее популярным способом последовательного извлечения из изображения рукописного слова символов марковской цепи в целях последующего распознавания является оконное сканирование изображения слева-направо с последовательной классификацией попадающих в окно графических элементов [15].

На рис. 1 показано рукописное слово, которое разбивается на четыре части по размеру сканирующего окна, обозначенного пунктирной линией. Стрелкой показано направление сканирования, а именно: слева-направо.

Рис. 1. Оконное сканирование

Размер окна на рисунке показан условно, на практике размер сканирующего окна необходимо выбирать в долях от среднего размера буквы; также, для увеличения эффективности процесса, «скроллинг» необходимо осуществлять с некоторым наложением.

Значительная вариативность в написании рукописных слов [16] и невозможность их ранжирования по количеству символов приводит к необходимости анализа изображения слова с использованием моделей, которые могут быть построены на основе слов, содержащих большее количество букв.

На рис. 2 показано слово, которое идентично предыдущему с разницей в одну последнюю букву. Для извлечения из него символов марковской цепи понадобится проанализировать ещё одно дополнительное «пятое» состояние.

Рис. 2. Более длинное слово

В случае формирования моделей рукописных слов для приведённых на рисунках выше примеров они будут различаться количеством состояний моделей, как показано на рис. 3 (четыре и пять состояний соответственно).

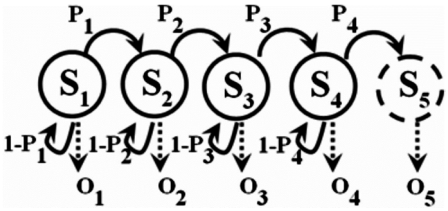

Рис. 3. Дополнительное пятое состояние

Проблема заключается в том, что модели, построенные на основе более длинных слов (в данном случае на основе слова «words»), имеют математическое преимущество перед моделями, построенными на основе слов более коротких (слово «word»). Математическое описание данного эффекта «поглощения» будет показано ниже.

Целью данной работы является демонстрация метода предотвращения ошибок в распознавании слов, отличающихся длиной окончаний.

Эффект «поглощения»

Ниже показана процедура расчёта вероятности соответствия распознаваемого изображения моделям с разницей в одно состояние. В формулах используются следующие обозначения:

$A\{a_{ij}\}$ – матрица вероятностей перехода из одного состояния в другое;

$B\{\ b(k)\ \}$ – матрица вероятностей генерирования наблюдаемых символов;

$i = \overline{1, N}$;

$j = \overline{1, M}$;

N – количество состояний $S_1, S_2, S_3 \ldots S_N$;

M – размер алфавита символов наблюдений (кол-во типов наблюдаемых символов);

$O_1, O_2, O_3 \ldots O_T$ – символы наблюдений;

p_i – вероятность того, что S_i – это начальное состояние модели;

$c(O|л)$ – вероятность того, что последовательность O порождена моделью л.

Формулы (1) и (2) показаны для расчёта вероятности для модели с четырьмя состояниями [15]:

$$б_1(i) = p_i b_i(O_1). \quad (1)$$

$$б_{t+1}(j) = \left[\sum_{i=1}^{4} б_t(i)\, a_{ij}\right] b_j(O_{(t+1)}). \quad (2)$$

Формулы (3) и (4) показывают дополнительные расчёты, выполняемые для пятого состояния [15]:

$$б_1(5) = p_5 b_5(O_1). \quad (3)$$

$$б_{t+1}(5) = \left[\sum_{i=1}^{5} б_5(i)\, a_{i5}\right] b_5(O_{(t+1)}) > 0. \quad (4)$$

Формулы ниже (5) демонстрируют то, что в рамках данной статьи называется «поглощением» – математическое превосходство моделей больших размерностей.

$$\begin{aligned} c_1(O|л_4) &= \sum_{i=1}^{4} б_T(i), \\ c_2(O|л_5) &= \sum_{i=1}^{4} б_T(i) + б_T(5), \\ c_2(O|л_5) &> c_1(O|л_4). \end{aligned} \quad (5)$$

Таким образом, вероятность того, что слово «word» будет распознано как слово «words», а не «word», выше в силу особенностей построения математических выражений.

Если слова небольшой длины можно отсортировать по геометрическим размерам с достаточной уверенностью, то для слов, состоящих из большого количества букв, это становится невозможным в силу того, что слова из девяти букв могут быть одной длины со словами из десяти и даже одиннадцати букв, даже будучи написанными одним почерком или синтезированными из отдельных частей [17].

Нормирование горизонтальных размеров

Для предотвращения «поглощения» авторами предлагается выполнять дополнительное нормирование изображений по ширине с некоторым определённым шагом. Нормировать необходимо не только изображения для распознавания, но и изображения для создания моделей. Равенство количества состояний в анализируемом изображении с размерностью модели гарантирует отсутствие проявления проблемы «взаимного поглощения».

Принцип нормирования с шагом «Step» показан на рис. 4.

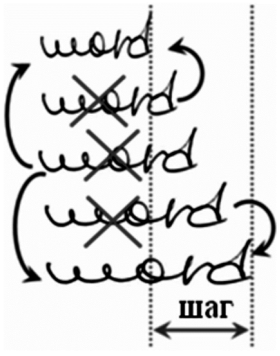

Рис. 4. Приведение длин графических изображений слов к нормированным размерам

Для оценки параметров нормирования нами был выполнен анализ 800 графических изображений рукописных слов, написанных двадцатью различными почерками.

Анализ графиков показал линейную зависимость среднеквадратичного отклонения длины слова от среднего значения его длины. Таким образом, выявлена необходимость увеличения области перекрытия линейных размеров распознаваемых слов несколькими размерами моделей слов. Это необходимо для того, чтобы в определении наилучшей модели участвовало наибольшее количество моделей слов, созданных из рукописных слов, написанных разными почерками.

Результаты анализа приведены на рис. 5.

Рис. 5. Зависимость среднеквадратичного отклонения длины слова от среднего значения в зависимости от количества букв в слове

Черная линия – линейная аппроксимация результатов эксперимента, который представлен линией синего цвета. Всего было проведено 50 экспериментов. Величина доверительного интервала составляла ±4.

Величина шага нормирования нами была выбрана исходя из средней ширины буквы в 40 пикселей. Авторами предлагается использовать коэффициент увеличения количества создаваемых моделей в зависимости от длины модели. Для слов, состоящих из десяти и менее букв, необходимо создавать две модели с определённым шагом нормирования. Для слов, содержащих до двадцати букв, необходимо создавать четыре модели, увеличивая таким образом диапазон, в котором может оказаться распознаваемое слово в два раза, и т.д.

При проведении эксперимента использовался подход, связанный со скелетизацией, аналогичный представленному в упомянутой ранее работе [15].

Сначала изображение текста мы записали в бинарную матрицу A_{ij}. Максимальные значения i и j определяются размером изображения в пикселях (разрешение изображения 300ppi). Затем определили толщину линий символов. Толщиной будет являться то количество подряд идущих пикселей по горизонтали (единиц матрицы A_{ij}), которое встречается наибольшее число раз. В случае совпадения значений для различных толщин итоговым значением толщины является минимальное. Полученное число уменьшается для гарантированного отсутствия вероятности образования разрывов.

Затем инициализируем нулями массив B_{ij}, который по размерам равен массиву A_{ij}. Далее к изображению мы применили маску с окружностью диаметром S, равным рассчитанной толщине линии. Окружность передвигается последовательно слева-направо и сверху-вниз с шагом в один пиксель. В случае полного перекрытия маски центральный пиксель окружности заносится в дополнительно организованный массив B_{ij}, в противном случае значение массива остаётся неизменным. В качестве признаков, используемых для описания признаков, использовались пиксели.

В итоге, мы получили массив, который описывает новое изображение, уже подвергнутое утонению, толщина которого определяется заложенной в алгоритм страховочной погрешностью. Размер погрешности был подобран экспериментально в два пикселя [18].

На этапе тестирования использовалось изображение уже выделенного слова. Использовалась база данных изображений русских слов [19]. В ней содержалось 6000 изображений, более 20 видов почерка.

Для извлечения из графического образа символов марковской цепи использовалось преобразование Хафа, которое в основном применяется для выравнивания рукописного текста относительно базовой линии [20]. В нашем случае преобразование использовалось для генерации множества прямых линий из участка изображения, попадающего в одну из пяти областей сканирующего окна.

Извлеченные линии сортируются по углу наклона (0^0, 45^0, 90^0 и 135^0) и по расположению в области сканирования (1, 2, 3, 4 и 5). Таким образом, получается двадцать видов символов. В терминах СММ – размер алфавита (M) равен двадцати. Из-за нестабильности в количестве генерируемых линий было применено двойное преобразование. Сначала находится функция распределения сгенерированных линий, а затем по этой функции строится последовательность символов фиксированной длины (в эксперименте использовалась последовательность длиной в двадцать символов). Функция распределения линий отражает не только факт заполнения определённой области, но и характер заполнения, в частности, угол наклона элемента изображения.

На рис. 6 приведен пример, используемый для тестирования алгоритма.

Рис. 6. Пример используемый для тестирования алгоритма

Проведённый эксперимент показал значительный эффект от нормирования. В среднем увеличение процента распознавания составило 7,7 процента по сравнению с алгоритмами (нейронными сетями), в которых не использовалась скелетизация.

Заключение

Таким образом, в статье продемонстрирована проблема «поглощения» моделей слов, созданных на основе слов, отличающихся окончаниями. Предложен подход с нормированием горизонтальных размеров с определённым шагом, который применим не только для задач, использующих марковское моделирование. Новым в работе является использование алгоритма нормировки в комбинации с СММ. Положительный эффект может быть получен также при применении для распознавания искусственных нейронных сетей [21].

Для подтверждения эффективности предложенного подхода проведён эксперимент, который показал наличие значительного эффекта от применения предлагаемого нормирования изображений рукописных слов по горизонтальному размеру.

Литература

1. CognitiveForms Paperless excellence [Электронный ресурс] - Режим доступа: http://cognitiveforms.com/ru/products_and_services/Cuneiform

2. Яковлев, С.С. Система распознавания движущихся объектов на базе искусственных нейронных сетей / С.С. Яковлев // ИТК НАНБ. – Минск, 2004. – С. 230-234.

3. Хювенен, Э. Мир Лиспа / Э. Хювенен, И. Сеппянен // в 2-х т. – М.: Мир, – 1990. – 318 с.

4. Bahlmann C. Online handwriting recognition with support vector machines - a kernel approach [Text] / C. Bahlmann, B. Haasdonk, H. Burkhardt // IEEE Transactions on Pattern Analysis and Machine Intelligence. – Vol. 26. – No. 3. – 2004. – P. 299-310.

5. Bentounsi H. Incremental support vector machines for handwritten Arabic character recognition [Text] / H. Bentounsi, M. Batouche // Proceedings of the International Conference on Information and Communication Technologies. – 2004. – P. 1764-1767.

6. Sanguansat P. Online Thai handwritten character recognition using hidden Markov models and support vector machines [Text] / P. Sanguansat, W. Asdornwised, S. Jitapunkul // Symposium on Communications and Information Technologies. – 2004. – Japan. – October 26-29. – 2004. – P. 492-497.

7. Bin Z. Support vector machine and its application in handwritten numeral recognition [Text] / Z. Bin, L. Yong, X. Shao-Wei // Proceedings of the 15th International Conference on Pattern Recognition. – 2000. – P. 720-723.

8. Shu H. On-Line Handwriting Recognition Using Hidden Markov Models [Text] / Han Shu // Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science. – February 1. – 1997.

9. Biadsy F. Online Arabic Handwriting Recognition Using Hidden Markov Models [Text] / Fadi Biadsy, Jihad El-Sana, Nizar Habash // The 10th international workshop on frontiers of handwriting recognition. – 2006.

10. Microsoft Windows 7 [Электронный ресурс] – Режим доступа: http://www.microsoft.com/rus/dino7/index.html.

11. Paragon software Многоязычный PenReader 9.0 [Электронный ресурс] – Режим доступа: http://www.penreader.com/.

12. Handwriting on the Go [Электронный ресурс] - Режим доступа: http://myscript.com/solutions/#mobility-section.

13. Horst Bunke Recognition of Cursive Roman Handwriting – Past, Present and Future / Horst Bunke // Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), – 2003, Volume 1, – pp. 448–459.

14. Norris D. Shortlist B: A Bayesian Model of Continuous Speech Recognition / D. Norris // Psychological Review, Vol. 115, No. 2, 2008 – pp. 357–395.

15. Мозговой А.А. Проблемы применения скрытых марковских моделей при распознавании рукописного текста / А.А. Мозговой // В мире научных открытий. 2013. – № 6. – С.186-198.

16. Sangeetha Devi S. Invariant and Zernike Based Offline Handwritten Character Recognition / S. Sangeetha Devi, Dr. T. Amitha // International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 5, May 2014, pp. 1950-1954.

17. Мозговой А.А. Методика синтеза словаря для задачи автоматического распознавания рукописных слов / А.А. Мозговой // Телекоммуникации. 2014. № 5. –С.3-4.

18. Мозговой А.А. Предварительная обработка изображений символов с целью улучшения качества последующей скелетизации (утонения) [Текст] / А.А. Мозговой // Вестник Воронежского института высоких технологий. – 2013. - № 10. – С. 156-160.

19. Мозговой А.А. Система распознавания рукописного текста с использованием математического аппарата скрытых марковских моделей [Текст] / А.А. Мозговой // Искусственный интеллект. Интеллектуальные системы ИИ-2013, материалы Международной научно-технической конференции (пос. Кацивели, АР Крым, 23 - 27 сентября 2013 года). – Донецьк:_ПШ_<Наука_осв_та>. – 2013. – С. 109-111.

20. Louloudis G. Text Line Detection in Unconstrained Handwritten Documents Using a BlockBased Hough Transform Approach / G. Louloudis, B. Gatos, C.Halatsis // Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, Volume 2. – pp. 599-603.

21. Vijay Laxmi Sahu Offline Handwritten Character Recognition Techniques using Neural Network / Vijay Laxmi Sahu, Babita Kubde // A Review IJSR Volume 2 Issue 1, January 2013 – pp. 87-94.