

УДК 004.352.243

И.Я. Львович, Я.Е. Львович, А.А. Мозговой,
А.П. Преображенский, О.Н. Чопоров

СОВЕРШЕНСТВОВАНИЕ АЛГОРИТМА РАСПОЗНАВАНИЯ РУКОПИСНЫХ ТЕКСТОВ НА ОСНОВЕ НОРМИРОВАНИЯ ИЗОБРАЖЕНИЯ

Воронежский институт высоких технологий

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 для уменьшения ошибок распознавания нормирование изображения.

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

1. ВВЕДЕНИЕ

Использование для распознавания целых рукописных слов марковского моделирования [1], а именно: лево-правой модели Бакиса [2] – 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 [3].

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

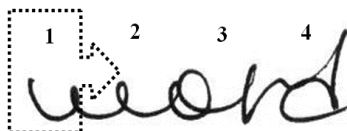


Рисунок 1 – Оконное сканирование

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 [4] 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 рисунке 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.

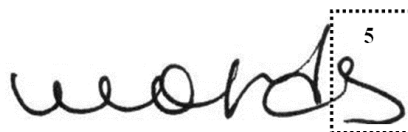


Рисунок 2 – Более длинное слово

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

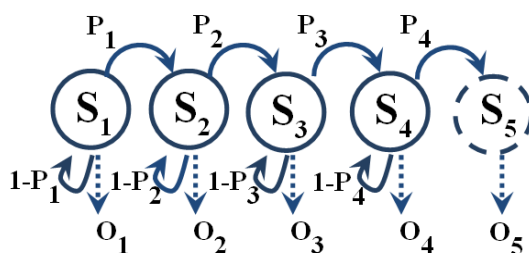


Рисунок 3 – Дополнительное пятое состояние

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.

2. 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;
- $B\{b_j(k)\}$ – матрица вероятностей генерирования наблюдаемых символов;
- $i = \overline{1, N}$;
- $j = \overline{1, M}$;
- N – количество состояний $S_1, S_2, S_3 \dots S_N$;
- M – размер алфавита символов наблюдений (кол-во типов наблюдаемых символов);
- $O_1, O_2, O_3 \dots O_T$ – символы наблюдений;
- p_i – вероятность того, что S_i – это начальное состояние модели;
- $c(O|L)$ – вероятность того, что последовательность O порождена моделью L .

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

$$\bar{\delta}_1(i) = p_i b_i(O_1). \quad (1)$$

$$\bar{\delta}_{t+1}(j) = \left[\sum_{i=1}^4 \bar{\delta}_t(i) a_{ij} \right] b_j(O_{(t+1)}). \quad (2)$$

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

$$\bar{\delta}_1(5) = p_5 b_5(O_1). \quad (3)$$

$$\bar{\delta}_{t+1}(5) = \left[\sum_{i=1}^5 \bar{\delta}_t(i) a_{i5} \right] b_5(O_{(t+1)}) > 0. \quad (4)$$

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

$$c_1(O|L_4) = \sum_{i=1}^4 \delta_T(i),$$
$$c_2(O|L_5) = \sum_{i=1}^4 \delta_T(i) + \delta_T(5), \quad (5)$$

$$c_2(O|L_5) > c_1(O|L_4).$$

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 [5].

3. 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 рисунке 4.

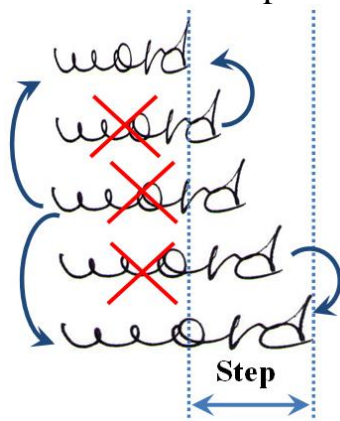


Рисунок 4 – Приведение длин графических изображений слов к нормированным размерам

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 рисунке 5.



Рисунок 5 – Зависимость среднеквадратичного отклонения длины слова от среднего значения в зависимости от количества букв в слове

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, аналогичный представленному в упомянутой ранее работе [3].

Для извлечения из графического образа символов марковской цепи использовалось преобразование Хафа, которое в основном применяется для выравнивания рукописного текста относительно базовой линии [6]. 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 скрытых марковских моделей – 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.

Проведённый эксперимент показал значительный эффект от нормирования. В среднем увеличение процента распознавания составило 7,7 процента.

4. 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. Положительный эффект может быть получен также при применении для распознавания искусственных нейронных сетей [7].

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

ЛИТЕРАТУРА

1. Horst Bunke Recognition of Cursive Roman Handwriting - Past, Present and Future // Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003, Volume 1, pp. 448–459.
2. Norris, D. Shortlist B: A Bayesian Model of Continuous Speech Recognition // Psychological Review, Vol. 115, No. 2, 2008 pp. 357–395.

3. Мозговой А.А. Проблемы применения скрытых марковских моделей при распознавании рукописного текста // В мире научных открытий. 2013. №6. С.186-198.
4. S. Sangeetha Devi, Dr. T. Amitha Invariant and Zernike Based Offline Handwritten Character Recognition // International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 5, May 2014, pp. 1950-1954.
5. Мозговой, А.А. Методика синтеза словаря для задачи автоматического распознавания рукописных слов // Телекоммуникации. 2014. №5. С.3-4.
6. Louloudis G., Gatos B., Halatsis C. Text Line Detection in Unconstrained Handwritten Documents Using a BlockBased Hough Transform Approach // Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, Volume 2. pp. 599-603.
7. Vijay Laxmi Sahu, Babita Kubde Offline Handwritten Character Recognition Techniques using Neural Network: A Review IJSR Volume 2 Issue 1, January 2013 pp. 87-94.

I.Y. Lvovich, Y.E.Lvovich, A.A. Mozgovoy,
A.P. Preobrazhensky, O.N. Choporov

**THE IMPROVEMENT OF THE ALGORITHM OF RECOGNITION OF
HANDWRITTEN TEXTS ON THE BASIS OF THE VALUATION
IMAGE**

Voronezh Institute of High Technologies

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 to reduce recognition errors, the normalization of the image.

Keywords: OCR, optical recognition, handwriting, HMM.

REFERENCES

1. Horst Bunke Recognition of Cursive Roman Handwriting - Past, Present and Future // Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003, Volume 1, pp. 448–459.
2. Norris, D. Shortlist B: A Bayesian Model of Continuous Speech Recognition // Psychological Review, Vol. 115, No. 2, 2008 pp. 357–395.

3. Mozgovoy A.A. Problemy primeneniya skrytykh markovskikh modeley pri raspoznavanii rukopisnogo teksta // V mire nauchnykh otkrytiy. 2013. No.6. pp.186-198.
4. S. Sangeetha Devi, Dr. T. Amitha Invariant and Zernike Based Offline Handwritten Character Recognition // International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 5, May 2014, pp. 1950-1954.
5. Mozgovoy, A.A. Metodika sinteza slovarya dlya zadachi avtomaticheskogo raspoznavaniya rukopisnykh slov // Telekommunikatsii. 2014. No.5. pp.3-4.
6. Louloudis G., Gatos B., Halatsis C. Text Line Detection in Unconstrained Handwritten Documents Using a BlockBased Hough Transform Approach // Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, Volume 2. pp. 599-603.
7. Vijay Laxmi Sahu, Babita Kubde Offline Handwritten Character Recognition Techniques using Neural Network: A Review IJSR Volume 2 Issue 1, January 2013 pp. 87-94.