УДК 528.854; 519.237

О.А. ТОДОРИКО, Г.А. ДОБРОВОЛЬСКИЙ
Запорожский национальный университет
М.Г. ДОБРОВОЛЬСКАЯ
Санкт-Петербургский государственный университет

ВЛИЯНИЕ МЕРЫ РАССТОЯНИЯ НА ТОЧНОСТЬ ОПРЕДЕЛЕНИЯ АВТОРСТВА СКАНИРОВАННЫХ РУКОПИСЕЙ

В работе исследуется изменение качества определения авторства рукописного текста с помощью предложенного авторами метода в зависимости от выбранной функции расстояния. Сравнение проводится для наиболее употребляемых метрик: Чебышева, городских кварталов и Евклида. Эксперименты показывают, что наилучшее качество достигается при использовании метрики Евклида.

Ключевые слова: оффлайновая аутентификация, аутентификация почерка, точность аутентификации, выделение графем, сравнение почерка, кластеризация, функция расстояния, метрика Евклида, метрика Чебышева, метрика городских кварталов.

О.О. ТОДОРІКО, Г.А. ДОБРОВОЛЬСЬКИЙ
Запорізький національний університет,
М.Г. ДОБРОВОЛЬСЬКА
Санкт-Петербурзький державний університет

ВПЛИВ МІРИ ВІДСТАНІ НА ТОЧНІСТЬ ВИЗНАЧЕННЯ АВТОРСТВА СКАНОВАНИХ РУКОПИСІВ

У статті досліджується, як змінюється якість визначення авторства рукопису за допомогою запропонованого авторами методу в залежності від вибраної функції відстані. Порівняння виконується для найуживаніших метрик: Чебишева, міських кварталів й Евкліда. Експерименти показують, що краща якість досягається при використанні метрики Евкліда.

Ключові слова: оффлайнова аутентифікація, аутентифікація почерку, точність аутентифікації, виділення графем, порівняння почерку, кластеризація, функція відстані, метрика Евкліда, метрика Чебишева, метрика міських кварталів.

O.A. TODORIKO, G.A. DOBROVOLSKY
Zaporozhye National University
M.G. DOBROVOLSKA
Saint Petersburg State University

INFLUENCE OF DISTANCE MEASURE ON ACCURACY OF WRITER IDENTIFICATION OF STATIC MANUSCRIPT IMAGES

The suggested before method of writer identification and verification of static manuscript images is elaborated. In the approach features of handwriting are extracted as distribution function of graphemes over a codebook. In this paper the identification accuracy is explored depending on distance measure used. The codebook is constructed in three ways using Euclidean, Manhattan or Chebyshev distances. Graphemes are curvilinear segments connecting points of interest like line ends, crossing and branching. They are extracted with consecutive binarization, line thinning, search of joints and lines between them, then loops are searched separately. Each grapheme is described by feature vector containing length of the segment, coordinates of it's start and end points, and robust shape description. The shape is described as histogram of tangent line directions. Then feature vectors extracted from all manuscripts ares divided into clusters which form a codebook and the handwriting is mapped into grapheme frequencies in each cluster. To compare two handwritings the corresponding frequencies of graphemes are compared using chi-square criteria. The set of feature vectors is divided into clusters using Euclidean, Manhattan or Chebyshev distances. Experiments show that the best handwriting distinction can be reached with Euclidean distance measure.

Keywords: offline authentication, handwriting authentication, graphemes, handwriting comparison, image binarization, thinning of lines, join points of the image, hierarchical clustering, Euclidean distance, Manhattan distance, Chebyshev distance.

Введение

Исследование почерка относится к поведенческим биометрическим методам, которое часто требуется для определения автора рукописного текста, например в судебной экспертизе, в исторических исследованиях, архивах. Однако большое количество рукописных документов требует развития средств автоматизации поиска, и это стимулирует развитие научных разработок в области определения авторства рукописей.

Постановка проблемы

Одним из успешных способов представления почерка является выделение локальных признаков сканированного текста, сопоставление каждому найденному признаку некоторого набора чисел, и кластеризация полученного множества наборов с помощью методов машинного обучения. Исследования, ориентированные на составление словаря локальных признаков почерка, чаще всего посвящены подбору подходящих признаков и их представлений. Методам кластеризации и мере расстояния между представлениями уделяется меньше внимания, хотя они тоже влияют на качество и скорость сравнения почерков. Мера расстояния сильно зависит от выбранных признаков и их представления, поэтому зачастую сложно определить метрику, приводящую к наилучшим результатам [1].

Как правило, оптимальную метрику подбирают экспериментальным путем, ориентируясь на наиболее известные функции расстояния, подходящие к выбранному набору признаков.

Предложенный ранее авторами статьи метод [2] ориентирован на статические изображения произвольных рукописных текстов неизвестного содержания. Он учитывает, что слова могут быть написаны практически без отрыва пера от бумаги. Текст такого рода является сложной, не повторяющейся системой пересекающихся кривых – это требует дополнительных усилий для выделения устойчивых признаков почерка.

Целью статьи исследование данного метода: влияния выбора функции расстояния между наборами числовых признаков на точность определения авторства.

Анализ публикаций

Большая часть методов статистической аутентификации до начала работы с изображением проводят его подготовку: очищение картинки от шума; скелетизация; бинаризация или преобразование к оттенкам серого; отбрасывание областей с малым количеством пикселей; преобразование к единому размеру, выделение значимой области, прочее.

После предобработки изображению и/или его части ставятся в соответствие числовой вектор признаков [1, 3]. Используемые признаки должны учитывать возможные отличия в содержании текста.

Самые успешные из последних работ использовали методы k-соседних сегментов, SURF, SIFT, гистограмма направлений контура [3].

Для проверки неизвестных образцов применяются методы машинного обучения. Классификаторы на основе дистанции [4, 5], например, методы k-ближайших соседей, k-средних, основанные на плотности точек в пространстве признаков [6, 7] относят неизвестный образец к одному из классов, вычисляя расстояние от него до каждого класса.

Для вычисления расстояния используются разные метрики [5]. Большинство исследователей применяет в алгоритмах определения авторства Евклидову метрику [8]. Её очевидной модификацией является добавление весов для каждого признака [9]. Также использовались: квадрат Евклидова расстояния [8], Манхеттенская дистанция или расстояние городских кварталов [8, 10], хи-квадрат и его модифицированная версия [8, 11, 12], расстояния Чебышева, Хемминга [8, 11], корреляционная метрика [13], дистанция Хаусдорфа [14], наибольшая общая подпоследовательность [15].

Успешность применения метрики для классификации зависит от природы вектора признаков, поэтому тяжело сделать однозначный вывод о преимуществе какой-либо формулы расстояния.

Для исследования разработанного авторами [2] способа сравнения почерков и выбора оптимальной функции расстояния необходимо решить задачи: предварительная обработка изображений, выделение пригодных для анализа почерка графем, подбор числовых характеристик, составление вектора признаков почерка с помощью методов машинного обучения. Экспериментальная проверка разработанного способа на известных тестовых наборах Firemaker и IAM и выбор функции расстояния по наилучшему среднеквадратическому отклонению.

Описание метода. Исходными данными для рассматриваемого метода являются сканированные изображения рукописного текста, который не должен содержать большого количества помарок, зачеркиваний и т.д. Прежде чем приступить к выделению и анализу графем, происходит предобработка рукописей: преобразование к оттенкам серого, проводится бинаризация и скелетизация изображения.

Разделение изображения на графемы начинается с выделения особых точек, в которых заканчивается 1, 3 или более кривых (исследуются ближайшая окрестность каждого черного пикселя - квадрат 3x3). Точка считается особой, если в ее окрестности находятся 1, 3 или 4 раздельные группы черных пикселей. Если в окрестности точки расположены 2 раздельные группы, значит это просто часть кривой. Все соединяющие особые точки отрезки кривых и есть графемы. Далее происходит поиск петель, которые не содержат особых точек, на данном шаге это единственные неучтенные пиксели. В результате весь исходный текст преобразован в набор кривых (рис. 1).

Полученные наборы графем достаточно велики, что позволяет пользоваться статистическими методами, однако затрудняет сравнение двух наборов.

Рисунок 1. а) выделение особых точек (обозначены кругами);
б) полученный после удаления особых точек набор графем.

Упрощения сравнения происходит за счет уменьшения размерности описания почерка за счет кластеризации. Для разбивки на кластеры необходимо каждому криволинейному отрезку поставить в соответствие набор числовых признаков, задается функция расстояния между ними. Признаками контура были выбраны: количество черных пикселей, координаты концов кривой и приближенная форма контура. Приближенная форма контура вычисляется в несколько этапов, сначала выбирается начальная точка - конец кривой, находящийся наиболее близко к началу координат. Для замкнутых кривых начальная точка выбирается искусственно. Далее составляется таблица значений, описывающая контур в параметрическом виде, например, $x(s)$ и $y(s)$, где s - длина пути вдоль контура от начальной до текущей точки. Кластеризация выполнялась последовательно в несколько итераций, каждая следующая - на результатах предыдущей. Такой способ выбран для ускорения, упрощения и улучшения качества работы метода.

Результатом кластеризации является разбиение пространства признаков на области. Для каждого почерка из базы было построено 4-х уровневое дерево кластеров. На первом уровне рассматривалась длина отрезка, на 2-м - координаты его начала, на 3-м - координаты его конца, на 4-м – свойства кривизны (количество интервалов убывания и возрастания для каждой координаты). Кластеризация выполнялась с помощью метода *k*-средних, который разбивал каждую группу исходных данных достаточного объема на несколько частей.

Неизвестный образец распределяется по тем же кластерам.

Описанный способ сравнения почерков проверялся на известных тестовых наборах IAM и Firemaker. Для контроля качества из каждого образца почерка 8 страниц использовалось для обучения программы и 2 страницы для тестирования. Для успешной работы метода необходимо, чтобы у почерка имелось 2000 и более графем.

Экспериментальная проверка. В процессе кластеризации происходит вычисление расстояния между числовыми наборами признаков. Для выбора оптимальной функции расстояния проведено вычисление отличий распределения для различных тестовых наборов при использовании расстояния Чебышева (таблица 1), городских кварталов (таблица 2) и Евклида (таблица 3).

Таблица 1.

Отличие распределений при выборе меры Чебышева

Образец	Анонимный почерк	Количество графем	Отличие распределений	Количество кластеров
000-learn	000-test	8847	4298,786908	967
000-learn_151	000-test_151	3288	759,7794986340	302
000-learn_152	000-test_152	3168	688,6389649	323
000-learn_153	000-test_153	3348	699,0343146	337
000-learn_154	000-test_154	3753	787,2682177	302
000-learn_345	000-test_345	2827	778,0566878	386

Таблица 2.

Отличие распределений при выборе меры городских кварталов

Образец	Анонимный почерк	Количество графем	Отличие распределений	Количество кластеров
000-learn	000-test	8847	4390,514929	974
000-learn_151	000-test_151	3288	626,283021	309
000-learn_152	000-test_152	3168	697,2099535	337
000-learn_153	000-test_153	3348	734,0755928	316
000-learn_154	000-test_154	3753	691,9304225	316
000-learn_345	000-test_345	2827	674,0058666	372

Таблица 3.

Отличие распределений при выборе меры Евклида

Образец	Анонимный почерк	Количество графем	Отличие распределений	Количество кластеров
000-learn	000-test	8847	4132,31545	981
000-learn_151	000-test_151	3288	783,8139628	302
000-learn_152	000-test_152	3168	584,6923247	288
000-learn_153	000-test_153	3348	562,5050126	288
000-learn_154	000-test_154	3753	781,3157632	344
000-learn_345	000-test_345	2827	656,6727103	358

В результате всех 18 проверок (по 6 для каждой функции расстояния) было успешно установлено авторство.

Для выбора оптимальной функции расстояния необходимо вычислить среднеквадратичное отклонение, данный показатель позволяет оценить, насколько значения во множестве могут отличаться от среднего значения. В идеальном случае отличие распределений между двумя рукописями одного автора стремится нулю, поэтому в качестве среднего значения выбирается значение 0. Результат представлен в таблице 4.

Среднеквадратическое отклонение вычисляется по формуле:

$$Sr = \sqrt{\frac{1}{n}\sum_{1}^{n}(x_i - \dot{x})^2},$$

где n –количество соответствий, x_i – отличие распределений , $\dot{x}$ – среднее значение.

Таблица 4.

Среднеквадратические отклонения.

Мера	Среднеквадратическое отклонение
Мера Чебышева	1881,70445
Мера городских кварталов	1898,546465
Мера Евклида	1797,699719

Выводы

В данной работе исследован метод аутентификации, основанный на способе выделения графем как криволинейных отрезков с границами в местах пересечения, слияния, ветвления и окончания линий. Это наиболее соответствует интуитивному представлению об элементарных навыках написания слов, что позволяет рассматривать графемы, как линии, написанные одним непрерывным подсознательным движением. Полученный набор графем анализируется методами машинного обучения, основанными на обработке числовых признаков каждой графемы, использовалось меньшее количество числовых характеристик каждой графемы, чем в аналогичных работах, что позволило сократить время на машинное обучение.

Оценка работоспособности метода проводилась на известных тестовых наборах данных IAM и Firemaker.

Экспериментальная проверка подтвердила качество работы метода, а так же по результатам эксперимента была выявлена оптимальная функция расстояния: выбор производился среди трех мер и лучшей оказалась мера Евклида.

Список использованной литературы

1. Sameh, M. Awaida State of the art in off-line writer identification of handwritten text and survey of writer identification of Arabic text [Electronic resource] / Sameh M. Awaida, Sabri A. Mahmoud // Educational Research and Reviews Vol.7(20). – 2012, P. 445-463, 25 . − Access mode: http://www.academicjournals.org/article/article1379684852_Awaida%20and%20Mahmoud.pdf
2. Тодорико, О.А. Метод выделения графем для сравнения почерка в сканированных рукописях [Текст] / О.А.Тодорико, Г.А. Добровольский, М.Г. Добровольская, //Вестник Херсонского национального технического университета. – Херсон: ХНТУ. – 2014. – № 3(50). – С. 174-178.
3. Jain, R. Combining Local Features for Offline Writer Identification, Frontiers in Handwriting Recognition (ICFHR) [Text] / R. Jain, D. Doermann // 14th International Conference on 1-4 Sept. – 2014. – P. 583-588,

doi: 10.1109/ICFHR.2014.103
4. Friedman, M. Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches [Text] / M. Friedman, A. Kandel. − World Sci. Publishing Company, 1999. − 329 p.
5. Дюран, Б. Кластерный анализ [Текст] / Дюран Б. и Оделл П. Пер. с англ. Е. З. Демиденко. Под ред. А.Я. Боярского. Предисловие А. Я. Боярского. − М., Статистика, 1977. − 128 с.
6. Ankerst, Mihael. OPTICS: Ordering Points To Identify the Clustering Structure [Text] / Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander // ACM SIGMOD international conference on Management of data. ACM Press. − 1999. − P. 49–60.
7. Ester, Martin. A density-based algorithm for discovering clusters in large spatial databases with noise [Text] / Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu // In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. − 1996. − P. 226–231.
8. Abdi, M. A. Novel Approach for Off-Line Arabic Writer Identification Based on Stroke Feature Combination [Text] / Abdi M, Khemakhem M, Ben-Abdallah H // 24th International Symposium on Computer Information Science, IEEE. − 2009. – P. 597-600.
9. Li, X. Writer Identification of Chinese Handwriting Using Grid Microstructure Feature [Text] / Li X, Ding X, In: Tistarelli M and Nixon M (Eds.) // Advances in Biometrics. 5558. − 2009. − P 1230-1239.
10. Srihari, S.Writer Verification of Arabic Handwriting [Text] / S. Srihari, G. Ball // The Eighth IAPR International Workshop on Document Analysis Systems. − 2008. − P. 28-34.
11. Bulacu M, Schomaker L (2007). Text-Independent Writer Identification and Verification Using Textural and Allographic Features [Text] / // IEEE Trans. Pattern Anal. Mach. Intell. 29(4): 701-717.
12. Brink, A. Towards Robust Writer Verification by Correcting Unnatural Slant [Text] / A. Brink, R. Niels, R. van Batenburg, C. van Den Heuvel, L. Schomaker // Pattern Recognit. Lett. − 2010. − 32(3) . – P. 449-457.
13. Zhang, B. Handwriting Pattern Matching and Retrieval with Binary Features [Text] / Zhang B // Ph.D. dissertation, Department of Computer Science and Engineering, State University of New York, Buffalo, NY. – 2003. – 172 p.
14. Schomaker, L. Sparse-Parametric Writer Identification Using Heterogeneous Feature Groups [Text] / L. Schomaker, M. Bulacu, M. van Erp // Proceedings of the International Conference on Image Processing. – 2003. – vol. 1. – P. 545-8.
15. Helli, B. A Writer Identification Method Based on XGabor and LCS[Text] / B. Helli, M. Moghaddam // IEICE Electron. Express. – 2009. – vol. 6(10). – P. 623-629.