УДК 528.854; 519.237

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

МЕТОД ВЫДЕЛЕНИЯ ГРАФЕМ ДЛЯ СРАВНЕНИЯ ПОЧЕРКА В СКАНИРОВАННЫХ РУКОПИСЯХ

Задача оффлайновой аутентификация личности по особенностям почерка в сканированной рукописи решается с помощью автоматического поиска и анализа графем – замкнутых петель и криволинейных отрезков, соединяющих особые точки рукописного текста. Предварительные оценки показали, что выбранный метод определения графем позволяет успешно определить автора рукописного документа.

Ключевые слова: оффлайновая аутентификация, аутентификация почерка, выделение графем, сравнение почерка, бинаризация изображения, утоньшение линий, особые точки изображения, иерархическая кластеризация.

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

GRAPHEMES EXTRACTION METHOD FOR HANDWRITING COMPARISON

Annotation

The paper considers offline automatic writer authentication of manuscript that is provided as a set of images. Presented an original feature extraction algorithm which can be applied to handwriting comparison. The distinction of handwritings is performed with extracting and analyzing a set of graphemes. Analysis of manuscript is carried out by successive image binarization, line thinning, search of joints, extracting separated with join points removal graphemes which are thin curve segments or loops. Then each grapheme is mapped to feature vector containing length of the line, coordinates of it's start and end points, and robust shape description. The shape is described with several indexes like number of intervals, where the horizontal coordinate increases when moving along the segment. The feature vectors are clustered using hierarchical K-means approach producing a tree which is considered as handwriting description. To check if the anonymous manuscript is written by a writer we split it into graphemes and distribute them over the cluster tree generated from the known samples of writer's handwriting. Then occurrences of anonymous graphemes are compared to frequencies in the cluster tree leafs using chi-square criteria. The preliminary numerical experiment corroborates our statement that the suggested grapheme and feature extraction method provides sufficient information to discriminate two handwritings.

Keywords: offline authentication, handwriting authentication, graphemes extraction, handwriting comparison, image binarization, thinning of lines, join points of the image, hierarchical clustering.

В наше время всё больше важной информации доступно в электронном виде. Но двоичный код безлик, его может написать кто угодно, и в важных случаях доверять непроверенной информации нельзя. Необходим надёжный способ подтверждения авторства. Используемая в настоящее время электронная цифровая подпись [1] надёжна лишь до тех пор, пока её владелец не потерял пароль и закрытый ключ, или пока злоумышленник их не украл [2]. Поэтому постоянно совершенствуются методы, позволяющие проверить авторство документов.

Ниже рассматривается аутентификация личности по особенностям почерка в сканированной рукописи. Эта задача называется оффлайновой (статической) аутентификацией и, в отличие от достаточно точно решенной онлайновой (динамической) аутентификации, не располагает динамическими параметрами, отражающими процесс написания текста: зависимостью координат пишущего острия от времени, силой его давления на поверхность и так далее [3]. Оффлайновая аутентификация имеет дело с изображениями рукописного текста, полученными с помощью сканера или фотоаппарата, исследует только статические свойства почерка и считается более сложной.

Целью исследования является составление и реализация алгоритма автоматического выявления признаков, которые могут применяться для сравнения двух образцов почерка, каждый из которых представлен в виде изображений произвольных страниц рукописи. Особо следует отметить, что аутентификация по почерку не эквивалентна распознаванию рукописного текста на изображениях, и распознавание не входит в задачи данной работы.

Для достижения поставленной цели следует решить следующие задачи: предварительная обработка изображений, выделение пригодных для анализа почерка фрагментов, подбор числовых характеристик для описания каждого фрагмента, выявление признаков почерка с помощью методов машинного обучения.

При этом необходимо принимать во внимание, что почерк может изменяться в зависимости от ручки, которой человек писал, настроения, состояния, бумаги, наличия на ней дополнительных пометок, водяных знаков.

Чаще всего при оффлайновой аутентификации проверяется рукописная подпись на чеках, счетах и других документах [4]. Для решения данной задачи можно применить методы сравнения изображений. Более сложным из-за различий сравниваемых текстов считается определение почерка на основе

произвольного рукописного текста [5].

Анализ публикаций. Большинство методов оффлайновой аутентификации перед началом работы с изображением проводят его подготовку: удаление шума при помощи фильтров; скелетизация (утоньшение); размытие и бинаризация или преобразование к оттенкам серого; отбрасывание областей с малым количеством пикселей; поворот подписи в горизонтальное положение; преобразование к единому размеру, выделение значимой области - описанного вокруг подписи прямоугольника.

На следующем этапе изображению и/или его части ставятся в соответствие числовой вектор признаков[5]. При этом признаки, используемые в оффлайновой аутентификации по произвольному рукописному тексту, должны учитывать возможные отличия в содержании текста.

Самое очевидное из локальных свойств почерка - это повторяющиеся мелкие детали рукописного текста, элементарные графемы. Использование графем впервые было применено в работе [6], и с тех пор было реализовано разными исследователями. В оригинальном методе границами графемы считаются минимумы на верхнем контуре рукописной строки. Соединение двух соседних элементарных графем даёт биграмму, трех - триграмму. Для выявления устойчивых отличительных признаков почерка графемы кластеризируются с помощью К-средних, карт Кохонена или других алгоритмов. Успешность распознавания составляла 93% для графем 1-го уровня, 95.45% для биграмм, 80% для триграмм.

Bar-Yosef, Beckman, Kedem, Dinstein в своей работе [7] выделяли из рукописного текста отдельные символы, каждый из которых, в свою очередь, описывался несколькими признаками: отношением площади доминирующего цвета к площади описанного выпуклого многоугольника, отношением полуосей описанного эллипса, свойствами кривизны, моментами распределения пикселей по площади описанного многоугольника.

В работе Siddiqi и Vincent [8] линии на изображении рукописного текста покрывалось небольшими квадратами, размер которых был сопоставим с толщиной линии, и полученные мелкие изображения разбивались на кластеры, которые и считались признаками почерка.

В других работах [5], учитывались свойства градиента, структуры, выпуклости. Изображение делилось на $n \times m$ частей с одинаковым количеством черных пикселей в каждой из n строк и в каждом из m столбцов, и для каждой ячейки вычислялся вектор признаков.

Из глобальных признаков использовались: распределение градиентов и его интегральные характеристики, вероятности появления оттенков серого, вычисленные на её основе "энергия" и корреляция, обратный разностный момент (inverse difference moment), энтропия. Для учета текстуры оценивались распределения направлений контуров, количество вертикальных, горизонтальных, положительных, отрицательных штрихов, средняя высота и средний наклон символов строки, количество внутренних контуров и внешних кривых, толщина штриха, среднее расстояние между словами. Для полностью автоматического извлечения признаков использовалось скользящее окно. В каждой позиции окна вычислялись: количество, центр масс, дисперсия черных пикселей; позиция и направление контура самого верхнего и нижнего черного пикселя; количество переходов от черного к белому; доля пикселей между верхним и нижним черными пикселями. В масштабах отдельной строки исследовалась проекция строки на горизонтальную линию. Также применялись такие преобразования изображений: фильтры Габора, вейвлет-преобразование, интегральное преобразование контуров. Некоторые успешные техники вычисления признаков, например, гистограмма угловой координаты, были независимо изобретены разными группами исследователей. Часто на основе распределения градиентов вычисляется наклон символов, хотя существуют работы, ставящие под сомнение ценность данного признака.

Для выделения устойчивых отличительных признаков почерка и проверки неизвестных образцов применяются методы машинного обучения. Классификаторы на основе минимальной дистанции с помощью метода k-ближайших соседей относят неизвестный образец к одному из классов, вычисляя расстояние от него до каждого класса. Для вычисления расстояния используются разные метрики: Евклидова метрика, расстояние городских кварталов, расстояние хи-квадрат, расстояния Чебышева, Хемминга, Минковского, Махаланобиса, Бхаттачария, Хаусдорфа и другие [9]. Успешность применения метрики для классификации очень сильно зависит от природы вектора признаков, поэтому трудно сделать однозначный вывод о преимуществе какой-либо формулы расстояния. Тем не менее, многие исследователи считают, что к более точным результатам приводит метрика хи-квадрат [5].

Для классификации также можно использовать бинарные вектора и меру сходства между ними. Согласно экспериментам, лучшие результаты обеспечивает расстояние корреляции [10].

Классификаторы, основанные на дистанции, не обнаруживают кластеры неправильной формы. В таких случаях лучше использовать байесовские сети, метод опорных векторов, скрытые Марковские модели [11], классификаторы, основанные на плотности точек в пространстве признаков [12, 13], нейронные сети Хемминга, Кохонена, нечеткие классификаторы.

Основная идея данной работы отличается от существующих аналогов и состоит в анализе замкнутых петель и криволинейных отрезков, соединяющих особые точки рукописного текста –

пересечения и окончания линий. Ниже будет показано, что такой выбор предоставляет достаточно информации для идентификации почерка.

Описание метода. Исходными данными для программной системы сравнения почерков являются изображения рукописного текста, отвечающие определенным требованиям. Для успешного учета размера символов разрешение рисунков должно быть одинаковым, количество страниц для успешного применения методов статистики должно быть достаточно большим. Сам рукописный текст не должен содержать большого количества помарок, зачеркиваний и т.д. Прежде чем приступить к выделению и анализу графем, проводится предварительная обработка изображения: преобразование к оттенкам серого, бинаризация и скелетизация.

Преобразование цветного изображения в оттенки серого в цветовых пространствах YUV и YIQ, используемое в PAL и NTSC вычисляется по известной формуле [14] :

$$Y'=0.299R+0.587G+0.114B,$$

где Y' - яркость, и R,G и B - параметры цвета в модели RGB.

Бинаризация состоит в отображении оттенков серого в черный или белый цвет, в зависимости от того превышает ли яркость заданный порог. Если средняя яркость фона на всем изображении одинакова, то порог существует и его можно вычислить, например, с помощью алгоритма Оцу [15]. Однако, если яркость фона в разных частях изображения слишком отличается, глобального порога не существует, и его нужно вычислять локально для каждой части изображения.

Скелетизация или утоньшение преобразует широкие следы пера в цепочки шириной один пиксель. Наиболее популярным является алгоритм Зонга-Суня [16].

Выделение графем начинается с определения особых точек, в которых заканчивается 1, 3 или более кривых. Для этого исследуются ближайшая окрестность каждого черного пикселя - квадрат 3x3, в которой, после скелетизации, может находиться от 0 до 4 черных пикселей, не считая центрального. Чтобы точка считалась особой, в ее окрестности должны находиться 1, 3 или 4 раздельные группы черных пикселей. Случай, когда в окрестности точки расположены 2 раздельные группы, означает, что это просто часть кривой. После нахождения всех особых точек находятся все соединяющие их отрезки кривых - графемы.

Последним этапом выделения графем является поиск петель, которые не содержат особых точек. На этом шаге петли легко находятся, так как они остаются единственными неучтенными пикселями на рисунке. Таким образом, исходный текст преобразуется в достаточно большой набор кривых. Для дальнейшего анализа его необходимо кластеризировать.

Для разбивки на кластеры нужно каждому криволинейному отрезку сопоставить набор числовых признаков и определить функции расстояния. При выборе набора признаков следует помнить, что кластеризация в пространстве большой размерности затруднена, и поэтому необходимо выбрать минимальное количество самых важных характеристик, которые позволят вычислить меру сходства между тонкими цепочками пикселей. В процессе кластеризации часто вычисляется расстояние между наборами признаков, поэтому функция расстояния должна быть максимально простой. Например, из-за сложных операций умножения и извлечения квадратного корня Евклидово расстояние будет работать медленнее, чем метрика Чебышева или расстояние городских кварталов. В данной работе функцией расстояния была выбрана метрика Чебышева.

С учетом названных выше требований, 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, например самая нижняя точка на левой вертикальной касательной. 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. Исследуя зависимости $x(s)$ и $y(s)$, можно найти количество и длину участков возрастания и убывания, количество и длину участков со значениями выше и ниже среднего.

Кластеризация выполнялась последовательно в несколько итераций, каждая следующая проводилась на результатах предыдущей. Такой способ был выбран для ускорения, упрощения и улучшения качества.

Результатом кластеризации является разбиение пространства признаков на области, содержащие некоторое количество графем. Такой подход позволяет использовать для дальнейшего сравнения почерка аппарат математической статистики, когда для каждого образца почерка строится распределение вероятности появления графемы в зависимости от номера кластера. Неизвестный образец распределяется по тем же кластерам и полученные частоты с помощью критерия Пирсона [17] сравниваются с уже известным распределением. Результатом сравнения будет уровень значимости - вероятность ошибочного

отбрасывания гипотезы о равенстве распределений.

Экспериментальная проверка. Описанный способ кластеризации тестировался на сканированных образцах почерка, студенческих конспектах. Для каждого образца выбиралось 10 страниц рукописного текста, которые, в зависимости от размера рукописных символов разделялись на 12-17 тыс. графем, которые, в свою очередь, группировались в кластеры.

В результате для каждого почерка было построено 4-х уровневое дерево кластеров. На первом уровне использовалась 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.

Для контроля качества из каждого образца почерка 8 страниц использовалось для обучения программы и 2 страницы для тестирования. Страницы, предназначенные для тестирования, обрабатывались аналогично обучающей выборке, но без выполнения кластеризации. В результате для каждого образца строилось дерево кластеров, которое содержало 200-300 “листьев”, по которым распределялись элементы тестовой выборки. Полученный таким образом закон распределения с помощью критерия хи-квадрат Пирсона сравнивался с законом распределения обучающей выборки.

Для проверки работоспособности предложенного алгоритма каждый из образцов сравнивался с каждым тестовым набором. Для образцов, представляющих разный почерк, вычисленный уровень значимости (вероятность ошибки при отбрасывании гипотезы о равенстве образцов) стремился к нулю. Для образцов, представляющих один и тот же почерк, найденный уровень значимости находился в пределах 0,4-0,8.

Проведенный эксперимент подтвердил, что выбранные способы выделения графем и их последующей кластеризации позволяют определять похожий почерк.

Выводы. Предложенный способ выделения графем в местах пересечения, слияния, ветвления и окончания линий больше соответствует интуитивному представлению об элементарных навыках написания слов, так как данные точки вероятнее всего будут местами начала или окончания движения пера. Таким образом, графемы состоят из линий, написанных одним непрерывным подсознательным движением. Этим данный способ существенно отличается от рассмотренных ранее методов выделения графем с помощью точек, ближайших к нижнему краю строки или накладывания искусственной сетки. В силу особенностей выделения графем, рассмотренный способ можно применить для обработки текстов, написанных с помощью иероглифов.

В дальнейшем планируется проверить гипотезу о том, что естественное выделение линий, написанных непрерывно, позволит сократить количество входных данных и соответственно увеличить скорость сравнения почерков.

Литература

1. Идентификация по почерку и динамике подписи [Электронный ресурс]. – Режим доступа: https://sites.google.com/site/biometriy/dinamiceskie-metody/identifikacia-po-pocerku-i-dinamike-podpisi
2. Mutton, Paul Half a million widely trusted websites vulnerable to Heartbleed bug. Netcraft) [Electronic resource]. − Access mode: http://news.netcraft.com/archives/2014/04/08/half-a-million-widely-trusted-websites-vulnerable-to-heartbleed-bug.html
3. Биометрическая аутентификация по динамическим характеристикам подписи [Электронный ресурс]. – Режим доступа: http://www.secuteck.ru/articles2/sys_ogr_dost/biometrich-autentifikac-po-dinamich-harakter-podpisi/
4. Das Rajdeep. A Comparative Study of Biometric Authentication Based on Handwritten Signatures [Electronic resource] / Rajdeep Das, Sangeeta Dhar, Sabarni Das, Saurav Dutta, Subra Mukherjee // IJRET: International Journal of Research in Engineering and Technology Volume: 02 Issue: 12. – 2013, pp. 28-35 . − Access mode: http://ijret.org/Volumes/V02/I12/IJRET_110212004.pdf
5. 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, pp. 445-463, 25 . − Access mode: http://www.academicjournals.org/article/article1379684852_Awaida%20and%20Mahmoud.pdf
6. Bensefia A. Writer identification by writer’s invariants [Electronic resource]/ Nosary A., Paquet T., Heutte L. // Eighth International Workshop on Frontiers in Handwriting Recognition. IEEE Computing Society, Ontario, Canada. – 2002, pp. 274-279. − Access mode: http://lheutte.free.fr/download/iwfhr02bensefia.pdf
7. Bar-Yosef I. Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents [Electronic resource]/ Bar-Yosef I., Beckman I., Kedem K., Dinstein I. − Access mode: http://www.ee.bgu.ac.il/~dinstein/Publications/IJDAR%20fulltext%20on-line%20March%202007.pdf

8. Siddiqi I. Combining Global and Local Features for Writer Identification [Electronic resource] / Siddiqi I., Vincent N. – 2008. − Access mode: http://www.cenparmi.concordia.ca/ICFHR2008/Proceedings/papers/cr1007.pdf
9. Дюран Б. Кластерный анализ / 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.
10. Zhang B. Handwriting Pattern Matching and Retrieval with Binary Features / B. Zhang // Ph.D. dissertation, Department of Computer Science and Engineering, State University of New York, Buffalo, NY. − 2003. − p. 172.
11. Srihari S. Comparison of Statistical Models for Writer Verification [Electronic resource] / Srihari S., Ball G. // Proceedings on Document Recognition and Retrieval XVI San Jose, CA, USA. − 2009. − Access mode: http://www.cedar.buffalo.edu/~srihari/papers/SPIE2009-Stat.pdf
12. Ankerst Mihael. OPTICS: Ordering Points To Identify the Clustering Structure / Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander // ACM SIGMOD international conference on Management of data. ACM Press. − 1999. − pp. 49–60.
13. Ester Martin. A density-based algorithm for discovering clusters in large spatial databases with noise / 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. − pp. 226–231.
14. Recommendation ITU-R BT.470-7, Conventional Analog Television Systems [Electronic resource]. − Access mode: http://www.itu.int/rec/R-REC-BT.470/en
15. Otsu, N. A threshold selection method from gray-level histograms / IEEE Trans. Sys., Man., Cyber. 9. − 1979. − pp.62-66.
16. Zhang T. Y. A Fast Parallel Algorithm for Thinning Digital Patterns [Electronic resource] / T. Y. Zhang, C. Y. Suen. − Access mode: http://www-prima.inrialpes.fr/perso/Tran/Draft/gateway.cfm.pdf
17. Корн Г. Справочник по математике для научных работников и инженеров / Г. Корн, Т. Корн. – М.: Наука, 1986. – 832 с.