Идентификация цифровых шрифтов на основе морфологической ширины*

Н.А. Ломов, Л.М. Местецкий
nikita-lomov@mail.ru|mestlm@mail.ru
Россия, Московский государственный университет имени М.В. Ломоносова

Рассматривается задача определения шрифта и его начертания на цифровых изображениях, таких как сканированные изображения и фотографии. Признаковым описанием шрифта служит совокупность интегральных дескрипторов ширины отдельных символов. Приведено описание эффективного вычисления таких дескрипторов на основе медиального представления формы. Для идентификации шрифта строится взвешенная композиция метрических классификаторов, связанных с отдельными буквами. На обширной базе изображений низкого разрешения, растеризованных программно, метод демонстрирует высокую надежность и высокую скорость работы. Приведены примеры работы метода на изображениях из реального мира.

Ключевые слова: *цифровой шрифт, морфологическая ширина, дисковое покрытие, медиальное представление, L_1-метрика, бустинг*

Digital Font Identification Based on Morphological Width*

N.A. Lomov, L.M. Mestetskiy
Russia, Lomonosov Moscow State University

We consider the problem of the recognition of the typeface and its style on the digital images, such as scanned images and photographs. Feature description of the font is a set of cumulative width descriptors of the individual characters. The paper presents the effective implementation of these descriptors calculation procedure based on the medial representation of the shape. To identify the font the weighted composition of metric classifiers associated with individual letters is constructed. On the extensive database of synthetic low-resolution images the method demonstrates high reliability and computational efficiency. Examples of how method works on the real-world images are provided.

Keywords: *digital font, morphological width, disk cover, medial representation, L_1 distance, boosting*

Введение

Современная типографика использует тысячи различных шрифтовых гарнитур для дизайна и верстки газет, журналов, книг, объявлений, рекламных проспектов и другой печатной и медийной продукции. Единой классификации шрифтов нет, поэтому во многих случаях выбор шрифта для печати документа выполняется на основе образцов. При таком выборе часто возникает задача идентификации шрифта по образцу текста. Само происхождение образца может быть различным. Это может быть страничка текста на бумаге, изображение на экране, фотография вывески или объявления. Например, дизайнер хочет напечатать рекламу, используя шрифт, который он увидел на конфетной коробке в магазине. Задача состоит в том, чтобы определить название шрифта по имеющемуся изображению. В быстрой правильной идентификации шрифта заинтересованы дизайнеры, типографы, издатели для правильного подбора адекватного шрифта при производстве различной печатной продукции. В решении этой задачи также заинтересованы правообладатели шрифтов, поскольку они хотят обеспечить контроль лицензионного использования своих шрифтов.

Идентификация шрифта в общем виде ставится как задача поиска следующим образом. Имеется эталонный набор большого числа шрифтов, включающий несколько тысяч различных гарнитур. Такой набор называется файлом. Имеется также образец печатного текста, называемый запросом. Необходимо определить наиболее близкий к запросу шрифт в файле.

Известно несколько сервисов, продоставляющих услуги по распознаванию шрифта на основе изображения отдельных знаков, а также по подбору аналогов этого шрифта в имеющихся библиотеках, например [4, 3]. Алгоритмы поиска и идентификации шрифтов, используемые в этих системах, не раскрываются, а результаты, которые они показывют на практике, выглядят весьма слабо. Кроме того, известные сервисы работают только с латиницей, для распознавания кириллицы подобных систем нет. Существующие методы распознавания шрифтов, основанные на текстурных [11] и типографских [12] признаках, рассчитаны на работу с небольшой базой, содержащей несколько десятков шрифтов. Работ, предполагающих взаимодействие с обширной (более 1000 шрифтов) базой

Работа выполнена при поддержке РФФИ, грант 14-01-00716.
Работа опубликована по гранту РФФИ №16-07-20482.

очень немного, последние из них основаны на формировании кодовой книги из векторов локальных признаков, в качестве которых может использоваться HOG [5] или SIFT [6]. Известен также метод, который по аналогии с методом собственных лиц можно назвать «методом собственных букв» [10]. При этом перспективы метода [6] в области создания пользовательского сервиса выглядят предпочительнее всего, так как он позволяет обновлять базу шрифтов без особых затрат и не требует разметки изображения.

В данной статье мы рассматриваем подход к решению этой задачи путем построения меры сходства шрифта-запроса с эталонными шрифтами из файла и выбора «ближайшего соседа». Целью работы является исследование меры сходства, основанной на так называемой морфологической ширине шрифта. Новизна предложенного метода заключается в использовании морфологических признаков, основанных на оригинальной математической модели.

Дисковое покрытие фигуры

В основе предлагаемого дескриптора шрифта лежит понятие площади дискового покрытия фигуры, введенное в [1]. Напомним основные определения.

Определение 1. Фигурой называется замкнутая область на плоскости, ограниченная конечным числом непересекающихся замкнутых жордановых кривых.

Определение 2. Пустым кругом называется круг, целиком лежащий в фигуре.

Определение 3. Дисковое r-покрытие фигуры – объединение всех пустых кругов радиуса r.

Дисковое r-покрытие можно неформально представить как часть фигуры, которую можно, не выходя за границы, закрасить кистью, оставляющей круглый след радиуса r.

Определение 4. r-площадь фигуры — площадь ее дискового r-покрытия.

Определение 5. Морфологическая ширина фигуры – это r-площадь фигуры как функция от r.

Будем обозначать эту функцию, определенную для фигуры X, как $S_X(r)$. Морфологическая ширина является невозрастающей функцией параметра r. Она инвариантна относительно параллельного переноса и поворота фигуры, а относительно подобия обладает следующим свойством. Пусть X^k – фигура, полученная из X в результате преобразования подобия с коэффициентом k. Тогда

$$S_{X^k}(r) = k^2 S_X\left(\frac{r}{k}\right). \quad (1)$$

Это свойство позволяет легко получить морфологическую ширину фигуры, если известна морфологическая ширина фигуры, ей подобной.

В качестве фигур мы будем рассматривать отдельные буквы шрифта или, как в случае, например, буквы Ы, составные части букв. Примеры дисковых r-покрытий латинской буквы A шрифта Times New Roman для разных значений r представлены на рис. 1.

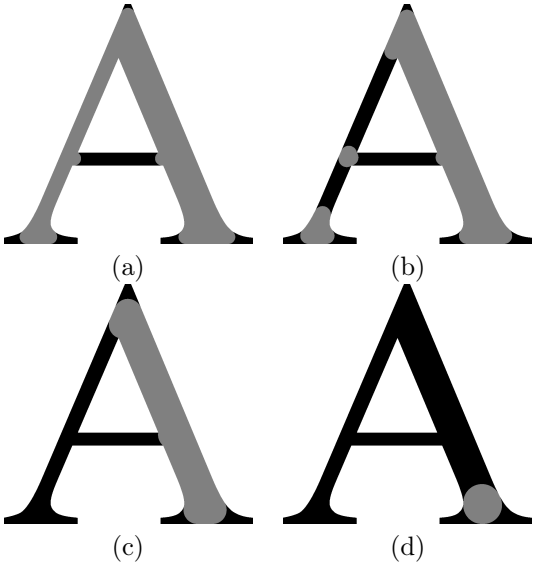


Рис. 1: Дисковое r-покрытие, если r составляет a) 3% b) 3,5% c) 6% d) 8% от высоты буквы

Площадь дискового покрытия можно найти по алгоритму, описанному в [1]. Форма при этом аппроксимируется многоугольной фигурой, для которой строится медиальное представление – скелет и радиальная функция. Дисковое r-покрытие многоугольной фигуры представляет собой область, ограниченную отрезками, лежащими на границе фигуры и дугами окружностей. В этой области можно выделить следующие части (рис. 2):

1. собственные области (полные или усеченные) ребер скелета – закрашены темно-серым ;
2. сектора вписанных кругов радиуса r – закрашены светло-серым;
3. линзы в пересечении этих секторов – закрашены синим.

Скелет фигуры выделен красным. Площадь покрытия можно найти как сумму площадей областей первых двух типов за вычетом площадей областей третьего типа. Все эти площади как функции от r имеют аналитическое представление, что позволяет рассчитать точную площадь r-покрытия на

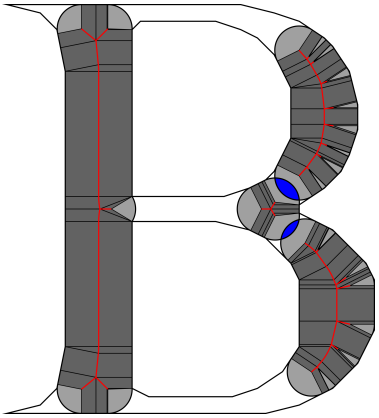


Рис. 2: Структура дискового покрытия многоугольной фигуры

основе высокоэффективной вычислительной процедуры.

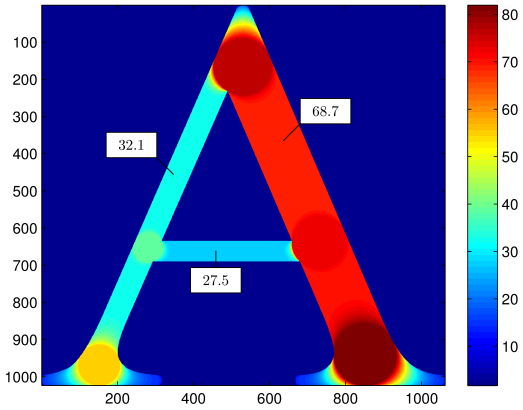


Рис. 3: Ширина в точках фигуры и значения в ее областях постоянства

Очевидно, что если точка фигуры входит в r-покрытие, то она также входит в любое покрытие в меньшим радиусом. Определим в точках фигуры функцию $T(p)$, равную максимальному радиусу покрытия, в которое входит точка p. Значение этой функции можно рассматривать как ширину фигуры в точке, так как она определяет размер максимального круга, который можно разместить в фигуре, покрыв точку p. Обратим внимание на то, что в случае, если буква имеет штрихи постоянной толщины, фигура имеет области, в которых значение функции $T(p)$ почти не изменяется. Визуализация такой функции с указанием ее стабильных значений представлена на рис. 3. Морфологическая ширина отражает распределение значений функции $T(p)$ и служит интегральным описанием ширины фигуры. На рис. 4 показан график морфологической ширины рассматриваемой буквы. Он является достаточно информативным, поскольку когда радиус становится равным толщине одного из штрихов буквы, на графике происходит резкий спад.

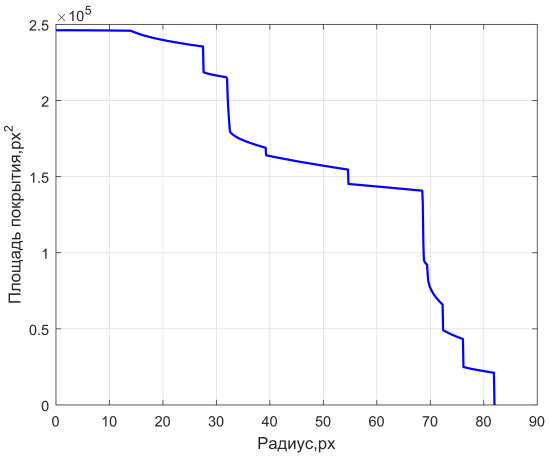


Рис. 4: Морфологическая ширина буквы, имеющей штрихи постоянной толщины

Сравнение шрифтов на основе диаграмм ширины

Как показывает рис. 5, ширинные характеристики различных шрифтов при фиксированной высоте символов очень сильно отличаются друг от друга. Однако символы одного и того же шрифта также имеют индивидуальные хорошо различимые портреты (рис. 6). Это не позволяет составить совокупную характеристику ширины шрифта, не зависящую от конкретного символа. Вместо этого мы будем рассматривать в качестве признакового описания шрифта набор дескрипторов ширины его отдельных символов. В качестве таких символов можно взять буквы латинского алфавита, что обеспечит универсальность метода, так как латиницу поддерживает подавляющее большинство шрифтов.

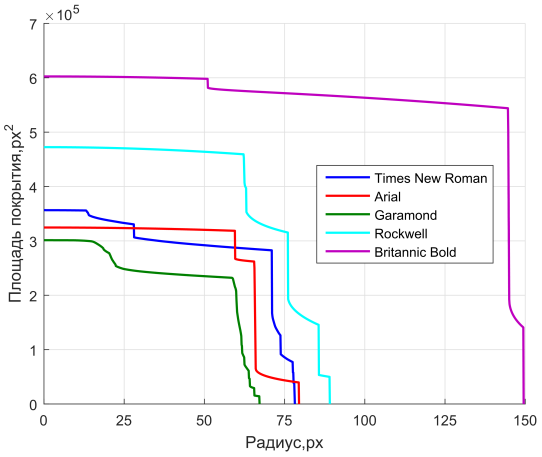


Рис. 5: Морфологическая ширина буквы H разных шрифтов

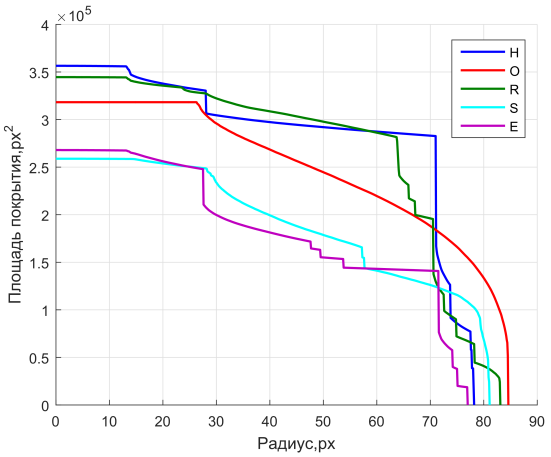


Рис. 6: Морфологическая ширина букв слова HORSE шрифта Times New Roman

Для построения эталонных дескрипторов шрифтов мы растеризуем выбранные символы в высоком разрешении (при этом размер шрифта должен быть одинаковым для всех символов) и вычислим значения r-площади при различных r, начиная от 0, с шагом h до тех пор, пока не будет получено нулевое значение. Размеры символов на входных растровых изображениях будут отличаться от эталонных, поэтому нам потребуется процедура нормализации. Для этого воспользуемся свойством 1 и для символа, размер которого в k раз меньше эталонного, вычислим значения r-площади с шагом $\frac{h}{k}$, а результаты умножим в k^2 раз. В результате мы фактически получим значения r-площади символа, увеличенного до эталонного размера, при тех же значениях аргумента, что у и эталонов.

Сравнение дескриптора символа на изображении с эталонами будем производить с помощью L_1-метрики. Так как мы имеем дело с дескрипторами, вообще говоря, разной длины, недостающую часть «короткого» дескриптора дополним нулями, поскольку площадь покрытия при этих значениях аргумента равна нулю. Для изображений, поступающих на вход алгоритма, должна быть произведена разметка – бинаризация, локализация символов, их определение и указание высоты заглавной буквы. Эти задачу можно решить при помощи методов оптического распознавания символов, широко освещенных в книгах [7, 9]. Обозначим код j-го символа на изображении c_j, его дескриптор ширины – q_j, эталонный дескриптор символа i-го шрифта с кодом j – p^i_j. Расстоянием от тестового образца $\mathbf{q} = \{q_1, \dots q_n\}$ до шрифта $\mathbf{p}^k$ будем считать взвешенную сумму расстояний до эталонов по отдельным символам:

$$d(\mathbf{p}^k, \mathbf{q}) = \sum_{j=1}^{n} \alpha_j d_1(p^k_{c_j}, q_j). \qquad (2)$$

В результате распознавания имеющиеся в базе шрифты ранжируются по мере сходства с представленным образцом, т.е. в порядке убывания расстояния, что позволяет получить как количественную (расстояния), так и качественную (номер в упорядоченном списке) оценку сходства.

Можно считать, что каждый символ задает элементарный классификатор a_j, использующий промежуточные оценки $f^k_j(\mathbf{q})$ – расстояния до k-го шрифта по j-му символу, а итоговая оценка $F^k(\mathbf{q})$ является взвешенной суммой таких оценок. Алгоритм выдает ответ в виде

$$a(\mathbf{q}) = \arg\min_k (F^k(\mathbf{q}))$$

и допускает ошибку на объекте $\mathbf{q}$, принадлежащем классу y, если

$$F^y(\mathbf{q}) > \min_{k \neq y} F^k(\mathbf{q}).$$

Поэтому пороговую функцию потерь можно аппроксимировать сверху следующим образом:

$$\begin{aligned}
[a(\mathbf{q}) \neq y] &\leqslant \exp\left(\beta\left(F^y(\mathbf{q}) - \min_{k\neq y} F^k(\mathbf{q})\right)\right) \leqslant \\
&\leqslant \exp\left(\beta \sum_{j=1}^{n} \alpha_{c_j}\left(f^y_{c_j}(\mathbf{q}) - \min_{k\neq y} h^k_{c_j}(\mathbf{q})\right)\right) \\
&\leqslant \exp\left(\beta \sum_{j=1}^{n} \alpha_{c_j} M_{c_j}(\mathbf{q})\right),
\end{aligned}$$

где $\beta > 0$ – параметр алгоритма, $M_{c_j}(\mathbf{q}) = f^y_{c_j}(\mathbf{q}) - \min_{k\neq y} h^k_{c_j}(\mathbf{q})$.

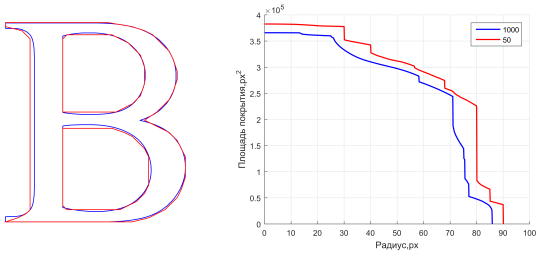
Это позволяет осуществить настройку композиции с помощью алгоритма AdaBoost [8], сформулированного для двух классов, в котором метки классов и ответы элементарных классификаторов $h(x)$ принадлежат множеству $\{-1, +1\}$. Достаточно произвести замены в алгоритме из вышеупомянутой статьи:

$$[h_t(x_i) \neq y_i] \Rightarrow [M_t(x_i) > 0],$$
$$h_t(x_i) y_i \Rightarrow -\beta M_t(x_i).$$

Эксперименты

Для создания базы эталонных дескрипторов были использованы 1884 шрифтовых начертания коллекции цифровых шрифтов компании Паратайп [2]. Отметим, что принадлежность начертаний к одному и тому же семейству с точки зрения нашего метода не играет никакой роли, и эти начертания

рассматриваются как полностью различные. Размер шрифта был выбран таким образом, чтобы высота заглавной буквы H составила 1000 пикселов, шаг радиуса при вычислении диаграмм был равен 0,1 пиксела. Создание базы эталонных диаграмм на компьютере с процессором Intel® Core i5™ и 6 ГБ оперативной памяти заняло 3,5 часа, причем почти все это время было затрачено на растеризацию букв большого размера, а построение скелетов и расчет диаграмм ширины заняли в сумме менее 8 минут.

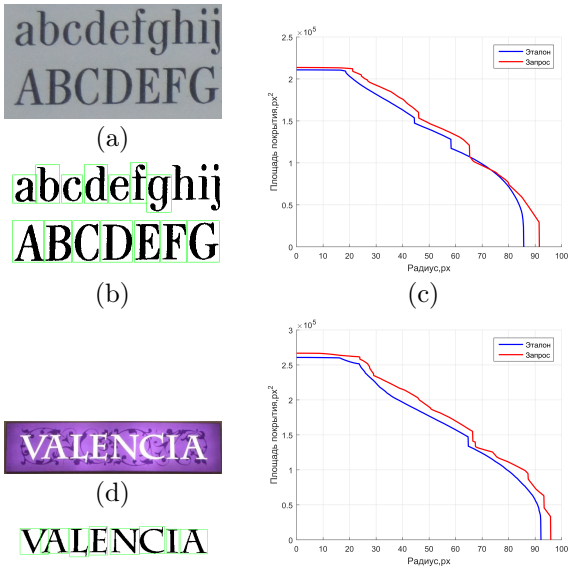


(a) (b)

Рис. 7: a) Масштабированные контуры букв высотой 1000 (синий) и 50 (красный) пикселов, б) их нормализованные диаграммы ширины

Для создания базы тестовых образцов символы этих же шрифтов были растеризованы в более низком разрешении так, чтобы высота буквы H составила 100 и 50 пикселов. Каждый образец содержал по одному изображению всех 52 букв латинского алфавита. Для этих символов также были построены диаграммы ширины с учетом нормализации. Примеры аппроксимации формы букв разного разрешения многоугольными фигурами, приведенными к единому масштабу, и их диаграмм ширины представлены на рис. 7. Веса классификаторов по отдельным буквам были получены на основе обучения по всему набору тестируемых шрифтов, параметр β был выбран равным 5×10^{-9} в силу того, что значения расстояний весьма велики. В итоговую композицию вошли 39 классификаторов, наилучшую дискриминирующую способность продемонстрировали символы S, C и Q. Результаты эксперимента показали, что уровень правильного распознавания шрифта при разрешении 100 составил 96%, а при разрешении 50 – более 91%. В ходе эксперимента были вычислены расстояния по 52 буквам между всеми парами из 1884 начертаний, что заняло 70 минут. Это означает, что время запроса – проверки одного начертания по базе эталонов – составило 2,25 секунды.

Для работы метода на реальных данных (первое изображение получено в результате сканирования, второе – в результате фотосъемки) проиллюстрирована на рис. 8. Разметка изображений была произведена вручную, высота заглавной буквы у первого образца составляет 70 пикселов, у второго – 63. В этих случаях метод успешно справляется с поставленной задачей и из 1884 кандидатов выбирает правильный шрифт.



(a) (b) (c) (d) (e) (f)

Рис. 8: Работа метода на реальных изображениях. a) d) Исходные изображения, b) e) их разметка, c) f) эталонная и тестовая диаграммы ширины буквы C

Выводы

Предложенный метод сравнения шрифтов использует в качестве основного дескриптора соотношение меду длиной и шириной штрихов в символах шрифта. Представление символов шрифта в виде непрерывных геометрических фигур и построение их дисковых покрытий позволяет получить количественное описание этого соотношения в виде функциональной зависимости площади дискового покрытия от размера покрывающих дисков. Экспериментыы показывают, что этот дескриптор является весьма информативным. Он позволяет учесть как индивидуальные особенности отдельных символов, так и особенности всего множества символов шрифта в целом. Практическая реализация предложенного метода использует эффективные алгоритмы вычисления непрерывного медиального представления для дискретных фигур, заданных бинарными растровыми изображениями. Высокая эффективность предложенного метода стала возможной благодаря использованию алгоритмов вычислительной геометрии.

Литература

[1] *Ломов Н.А., Местецкий Л.М.* Площадь покрытия многоугольной фигуры – дескриптор формы изображения – Компьютерная оптика (в печати).

[2] *Якупов Э., Петрова И., Фридман Г., Корольков а А., Левин Б.* 2008–2014 – ПараТайп. Цифровые шрифты – Москва, ПараТайп, 2015.

[3] `www.identifont.com` – Identifont – 2000–2016.

[4] `www.myfonts.com/WhatTheFont` – WhatTheFont! – 1999–2016.

[5] *Bui T., Collomosse J.* Font finder: Visual recognition of typeface in printed documents – Conference on Image Processing (ICIP), 2015 – pp. 3926-3930.

[6] *Chen G., Yang J., Jin H., Brandt J., Shechtman E., Agarwala A., Han T.* Large-Scale Visual Font Recognition – Conference on Computer Vision and Pattern Recognition, 2014. – pp. 3598–3605.

[7] *Cheriet M., Kharma N., Liu C.-L., Suen C.* Character Recognition Systems – A Guide for Students and Practitioners – John Wiley & Sons, Inc., Hoboken, NJ, 2007.

[8] *Freund Y., Shapire R. E.* A Short Introduction to Boosting – Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999 – pp. 1401–1406.

[9] *Rice S., Nagy G., Nartker T.* Optical Character Recognition: An illustrated guide to the frontier – Kluwer Academic Publishers, 1999.

[10] *Solli M., Lenz R.* FyFont: Find-your-Font in Large Font Databases – Proceedings of the 15th Scandinavian conference on Image analysis, 2007. – pp. 432–441.

[11] *Zhu Y., Tan T. Wang Y.* Font recognition based on global texture analysis – IEEE Trans. Pattern Analysis and Machine Intelligence, 2001. – vol. 10, no. 23 – pp. 1192-1200.

[12] *Zramdini A., Ingold R.* Optical font recognition using typographical features – IEEE Trans. Pattern Analysis and Machine Intelligence, 1998. – vol. 8, no. 20 – pp. 877-882.