Алгоритмы распознавания шрифтов в печатных документах

О.А. Славин

Аннотация. В статье описаны алгоритмы построения шрифтов на основе результатов распознавания печатных символов, используемые в механизме адаптивного распознавания документа. Определяется понятие шрифта, позволяющие построить алгоритмы распознавания шрифтов, с помощью которых производится дальнейшая кластеризация образов символов. Рассмотрены задачи, в решении которых может использоваться механизм поиска шрифтов: повторная сегментация границ символов, упаковка символов, подбор шрифта для отображения символов на экране.

Ключевые слова: шрифт, распознавания, кластеризация, адаптация.

1. Адаптивное распознавание и кластеризация

Механизм адаптивного распознавания отсканированных образов текстовых документов, описанный в [1,2], состоит из четырех основных этапов.

На первом этапе (первом проходе распознавания) происходит распознавание текста одним из шрифтонезависимых методов. В результате работы первого этапа для каждого распознаваемого образа указывается класс, к которому принадлежит образ, и дается некоторая оценка качества распознавания, то есть надежность принадлежности к выбранному классу.

На втором этапе (этапе кластеризации) происходит анализ результатов первого прохода распознавания. Среди надежно распознанных образов каждого класса проводится кластеризация.

На третьем этапе (этапе самообучения) полученные кластеры анализируются. Проводится поиск использованных шрифтов, отбор кластеров, обладающих наиболее благоприятными характеристиками – большей мощностью, лучшей надежностью распознавания составляющих объектов. Для каждого отобранного кластера строится эталон, наименее отличающийся от элементов кластера.

На четвертом этапе (*втором проходе распознавания*) проводится перераспознавание. Образы символов, ненадежно распознанные на первом этапе, проходят дополнительную проверку с помощью построенных на третьем этапе эталонов. В результате такой проверки некоторые образы символов могут быть отнесены к иным классам, у объектов могут быть изменены оценки надежности распознавания, может быть изменена сегментация некоторых слов.

Для успешной работы адаптивного распознавания оказывается плодотворной идея классификации символов с точки зрения принадлежности к шрифтам, использованным при печати документа. В настоящей работе рассматриваются понятия, алгоритмы поиска и приложения, относящиеся к распознаванию шрифтов на всех этапах адаптивного распознавания.

2. Атрибуты распознанных образов символов

В текстовом документе содержатся не просто различные символы, но слова, напечатанные одним или несколькими шрифтами. Различные символы, напечатанные одним шрифтом, имеют немало общего в своих характеристиках: согласованные размеры символов,

общие атрибуты (такие, как курсивность, жирность, наличие серифов), согласованное расположение на странице и т.п.

Символы одного шрифта обладают общими характеристиками, среди которых мы выделяем следующие:

- алфавит, то есть перечень кодов, соответствующих графическим образам символов;
- кегль, соответствующий расстоянию между базовыми линиями (алгоритмы построения базовых линий приведены в работе [3]), определяющий размеры образов символов между 2-ой и 3-й базовыми линиями (Рис. 1). Размеры остальных типов образов символов определяются пропорционально;
- признак моноширинности шрифта, определяющий близость ширин образов некоторых символов;
- признаки графем (начертаний образов символов):
  - признак курсива;
  - признак жирности;
  - признак наличия серифов;
  - признак подчеркивания;
  - признак зачеркивания;
  - признак особого начертания образа символа (*графема*);
  - код гарнитуры (полиграфического шрифта), которой может принадлежать данная графема;
- признаки однородности:
  - принадлежность слову;
  - принадлежность ячейке таблицы;
  - принадлежность строке;
  - принадлежность абзацу.

Значение кегля и признаки графем могут быть определены не для всех образов символов некоторого алфавита. Это объясняется как объективными причинам, так и алгоритмами определения этих характеристик, описанных в работе [4]. Примеры особенностей характеристик с точки зрения классификации символов приведены в Табл. 1.

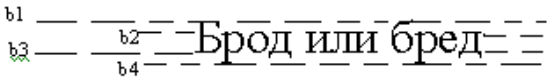


Рис. 1. Схема базовых линий

b1 - прямая, на которой начинаются заглавные буквы
b2 - прямая, на которой начинаются строчные буквы (кроме «б» и «ф»)
b3 - основная база, на которую опираются большинство букв, кроме опущенных («у», «р», «ф») и полуопущенных («Д», «Ц», «Щ»)
b4 – база, на которую опираются опущенные буквы

Для нахождения кода гарнитуры может применяться следующий способ. Для нескольких гарнитур $G=\{G_1, \ldots, G_N\}$, представленных последовательностями образов $P_1,\ldots,P_N$ с известными заранее кодами символов и значениями атрибутов, рассмотренных выше, рассчитываются таблицы эталонов для N признаковых алгоритмов сравнения с эталонами. В качестве таковых могут выступать методы нейронных сетей или полиномов, описанные в работе [5]. Для образа символа S производится анализ последовательности значений

$$W(S, \mathfrak{R}_1), \ldots W(S, \mathfrak{R}_N),$$

где $\mathfrak{R}_i$ – признаковый метод, обученный на последовательности образов, соответствующих i-ой гарнитуре,

$W(S, \mathfrak{R}_i)$ – оценка распознавания образа S с помощью метода $\mathfrak{R}_i$.

Табл. 1

Характеристика	Определяют	Не дают определить	Комментарий
Кегль	кенгшзхъываполжэячсмитьбю	йцущфрд	Символы, не размещенные между 2 и 3 базовыми линиями, не определяют кегль
Курсив	ЙЦЕКНГШЩЫПРДЖЧМТЬБЮЯ	*уегзхваоэяб*	Символы, образы которых не обладают вертикальными элементами, не определяют признак курсива
Жирность		**НТПРО**	Образы символов из полужирных полиграфических шрифтов не всегда могут быть отделены от нежирных при использовании выбранной функции расстояния
Сериф	ЙЦКНГШЩХЫПР	О89ЭС	Существуют символы, в образах которых отсутствуют горизонтальные серифы

В зависимости от особенности реализации методов $\Re_i$ возможно упрощенное ускоренное распознавание образа символа S с известным кодом символа.

Другой подход состоит в применении нейронных сетей или других алгоритмов, например, полиномов [5], обученных на совокупности $P_1+\ldots+P_N$ всех последовательностей образов символов так, что алфавит выходов нейронной сети включает код гарнитуры: $A' = A \otimes G$,

где A – алфавит обучения,

G – множество гарнитур.

В этом подходе код гарнитуры соответствует выходу, породившему максимальную оценку.

В обоих подходах существенным является требование к методам, состоящее в монотонности оценок надежности распознавания. Согласно результатам экспериментов, приведенных в [5], наибольшего доверия заслуживают алгоритмы нейронных сетей и распознавания полиномами.

3. Определение шрифта

В настоящей работе мы используем определение понятия *шрифт*, базирующееся на наборе образов символов $\overline{S}$, полученном распознаванием образа страницы.

О каждом распознанном на первом проходе символе $S \in \overline{S}$ имеется некоторая информация, содержащая:

- коллекцию альтернатив распознавания, состоящую из кодов символа и оценок надежности $\{(C_1,w_1), \ldots ,(C_n,w_n)\}$, код C_1, соответствующей наиболее надежной оценке w_1, будем называть кодом символа S. При этом код C_i, принадлежащий определенному алфавиту распознавания, может также определять графему;
- координаты (x_1,y_1,x_2,y_2) прямоугольника, охватывающего образ символа;
- базовые линии b_1, b_2, b_3, b_4, определяющие кегль $K(S)= b_3- b_2$;
- признаки графем, то есть признаки наличия серифов $P_s(S)$, жирности $P_b(S)$, наклона $P_i(S)$ (далее через P_λ будем обозначать один из признаков, P_s, P_b, P_i), а также код гарнитуры $G(S)$.

Также считается известным отображение

$$\wp: \overline{S} \times \overline{S} \to (0,1)$$

определяющее принадлежность пары символов к одному слову и иной группе однородности.

Мы считаем, что механизм распознавания на первом проходе обладает возможностью распознавания характеристик шрифта для образа каждого символа. При этом формируется оценка надежности распознавания для каждой характеристики. Считаем также, что механизм распознавания может формировать отказ от распознавания какой-либо характеристики образа символа, а также отказ от распознавания какой-либо характеристики для всех символов. Например, мы будем в основном рассматривать случай, когда невозможно распознавание кода гарнитуры, к которому принадлежит образ символа.

Случаю невозможности определения значения того или иного признака соответствует значение θ, отличное от области значений всех признаков. Для случая невозможности распознавания кода шрифта также задаем $G(S) = \theta$.

Шрифтом ϕ мы будем называть такую совокупность распознанных символов, в которой для любых двух символов S_A и S_B, выполнены условия:

- значения кегля символов одинаковы $K(S_A) = K(S_B)$;
- признаки графем одинаковы $P_\lambda(S_A) = P_\lambda(S_B)$,

  или $P_\lambda(S_A) = \theta$, но при этом $\exists S_X \in \phi$: $P_\lambda(S_X) \neq \theta$, $\wp(S_A, S_X) = 1$,

  или $P_\lambda(S_B) = \theta$, но при этом $\exists S_X \in \phi$: $P_\lambda(S_X) \neq \theta$, $\wp(S_B, S_X) = 1$;
- коды гарнитуры одинаковы $G(S_A) = G(S_B)$,

  или $G(S_A) = \theta$, но при этом $\exists S_X \in \phi$: $G(S_X) \neq \theta$, $\wp(S_A, S_X) = 1$,

  или $G(S_B) = \theta$, но при этом $\exists S_X \in \phi$: $G(S_X) \neq \theta$, $\wp(S_B, S_X) = 1$.

Таким образом, шрифт является совокупностью символов одинакового кегля и одинакового начертания, например, Times Roman и ***Times Roman Bold Italic*** являются различными шрифтами с одной гарнитурой.

Замечание 1: Согласно определению шрифта, символы, составляющие шрифт, содержат не более одного кегля и не более одного каждого из признаков графем.

Замечание 1 позволяет дать следующее определение:

Характеристиками шрифта ϕ будем называть:

- кегль $K(\phi)$ составляющих его символов;
- признаки графем $P_\lambda(\phi)$ составляющих его символов;
- код гарнитуры $G(\phi)$составляющих его символов.

Признаки графем и код гарнитуры шрифта ϕ могут быть неизвестны: $P_\lambda(\phi)=\theta$ или $G(\phi)=\theta$.

Определение шрифта, данное выше, является конструктивным, то есть позволяет на его основе строить алгоритмы поиска шрифтов на основе результатов распознавания первого прохода.

Простейший алгоритм поиска шрифтов состоит из трех шагов:

ШАГ 1. Выбираем символ S_1, у которого определен кегль $K(S_1)$. Этот символ порождает первый из шрифтов ϕ_1.

ШАГ 2. В процессе последовательного перебора всех символов, каждый из символов S, у которого определен кегль $K(S)$, либо присоединяется к одному из уже существующих шрифтов ϕ_k из множества шрифтов $\bar{\phi}=\{\phi_1 \cdots \phi_{m(\bar{\phi})}\}$ (если $K(S)=K(\phi_k)$ и $P_\lambda(S)=P_\lambda(\phi_1)$, $1\le k\le m(\bar{\phi})$), либо порождает новый шрифт $\phi_{m(\bar{\phi})+1}$.

ШАГ 3. К построенному на шагах 1 и 2 множеству шрифтов $\bar{\phi}$ добавляются символы, у которых не определен кегль, признаки или гарнитура. Осуществляется последовательный перебор групп однородности, в каждой из которых содержатся

- либо символы, принадлежащие одному единственному шрифту $\bar{\phi}_k \in \phi$,
- либо символы, у которых неизвестен кегль или код гарнитуры или отсутствуют признаки графем.

Символы с неизвестным кеглем, кодом гарнитуры или признаками графем, добавляются к шрифту $\bar{\phi}_k$.

В силу определения шрифта шаги описанного алгоритма не зависят от выбора первого символа, от способа перебора символов на шаге 2 и от способа перебора групп однородности на шаге 3.

Очевидно, что часть распознанных символов не попадет ни один из шрифтов, во-первых, из-за принципиальной невозможности определения характеристик некоторых символов, и, во-вторых, из-за ошибок распознавания характеристик символов. Поэтому мы предлагаем поиск шрифтов дополнить механизмами кластеризации.

4. Поиск шрифтов с помощью кластеризации

Кластеризация образов символов призвана классифицировать массивы информации о распознанных символах и преобразовать в группы для последующих этапов распознавания.

Перед кластеризацией каждый из бинарных образов $R(m,n)=\|r_{ij}\|$, где $i=\overline{1,m}$, $j=\overline{1,n}$, $r_{ij}\in\{0,1\}$, в котором 1 соответствует черной точке, а 0 - белой, подвергается нормализации размера и центрированию. Нормализованный и центрированный растр $\tilde{R}(M,N)=\|\tilde{r}_{ij}\|$ определяется следующим образом

$\tilde{r}_{(i-s_1),(j-s_2)}=r_{ij}$, если $s_1\le i<s_1+m-1$ и $s_2\le j<s_2+n-1$,

$\tilde{r}_{ij}=0$, если $i<s_1$, или $i\ge s_1+m$, или $j<s_2$, или $j\ge s_2+n$,

$$s_1=[(M-m)/2], s_2=[(N-n)/2],$$

причем размеры M и N являются общими для всех кластеров.

Пусть в пространстве нормализованных по размеру и центрированных образов задана функция расстояния, являющаяся *симметрикой*, то есть удовлетворяющая следующим условиям:

$d(x,y)\ge 0$ для $\forall x, \forall y$

$d(x,y)= d(y, x)$ для $\forall x, \forall y$

$d(x,x)=0$ для $\forall x$.

Опишем способ построения функции симметрики d. Рассмотрим единичную окрестность растра $R(m,n)=\|r_{ij}\|$, где $i=\overline{1,m}$, $j=\overline{1,n}$, $r_{ij}\in\{0,1\}$, то есть $N^{(1)}(R)(m,n)=\| N_{ij}^{(1)}(R)\|$,

где $N_{ij}^{(1)}(R)=\max(r_{(i-1)(j-1)}, r_{(i-1)j}, r_{(i-1)(j+1)}, r_{i(j-1)}, r_{ij}, r_{i(j+1)}, r_{(i+1)(j-1)}, r_{(i+1)j}, r_{(i+1)(j+1)})$, если $0<i<m1$ и $0<j<n-1$,

$N_{ij}^{(1)}(R)=r_{ij}$ если $i=0$, или $i=m-1$, или $j=0$, или $j=n-1$.

В качестве расстояния d между двумя бинарными растрами $A=\|a_{ij}\|$ и $B=\|b_{ij}\|$ берется

сумма числа точек первого растра A, выходящих за единичную окрестность второго растра $N^{(1)}(B)=\| b_{ij}^{(1)} \|$, и числа точек второго растра B, выходящих за единичную окрестность первого растра $N^{(1)}(A)= \| a_{ij}^{(1)} \|$:

$$d(A,B)= \sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij} \bullet \overline{b}_{ij}^{(1)} + b_{ij} \bullet \overline{a}_{ij}^{(1)}) .$$

Если размер символов очень велик, то радиус окрестности необходимо увеличить. Однако при стандартных разрешениях 200-400 точек на дюйм единичная окрестность представляется оптимальной с точки зрения алгоритма распознавания символов, основанного на результатах кластеризации.

Для кластеризации образов символов применялся метод *цепной развертки* [2], в результате работы которого исходное множество разделяется на несколько кластеров таким образом, чтобы цепное расстояние между любыми объектами, входящими в разные кластеры, было больше заданного порога r_0, а для любых объектов из одного кластера цепное расстояние было не больше r_0.

Результатом кластеризации является набор кластеров символов $Cl_1, \ldots, Cl_L$, каждый из которых состоит из набора одноименных элементов (символов) $Cl_i=\{S_{1i}, \ldots, S_{ki}\}$ со сходным начертанием. Каждому из символов S соответствует нормализованный и центрированный образ $\widetilde{R}(S)$ (размеры M и N у всех растров одинаковы). Каждый из символов S обладает следующим набором характеристик $\chi(S)$:

- код символа $C(S)$, принадлежащий алфавиту распознавания A и являющийся кодом альтернативы распознавания с наибольшей оценкой надежности;
- кегль символа $K(S)$;
- признаки графем $P_\lambda(S)$ (признаки жирности, серифности, наклона);
- код гарнитуры $G(S)$;
- код шрифта $F(S)$, который может быть определен с помощью алгоритма, описанного в разделе 3.

При этом код символа, признак графем, код гарнитуры и код шрифта могут быть не определены: $\chi(S) = \theta$.

Также считается известным отображение $\wp$, определяющее принадлежность пары символов к одной группе однородности.

В силу равенства размеров всех элементов символов $\{S_1, \ldots, S_k\}$, составляющих кластер Cl, определим *расширенный образ* символов кластера как сумму образов

$$Cover(Cl) = \| r_{ij}(Cl) \| = \sum_{s=1}^{k} \vec{r}_{ij}^{(s)}$$

где $\vec{r}_{ij}^{(s)}$ – точка образа символа S_s, возможно, подвергнутого сдвигу на 1 пиксель по вертикали или горизонтали (вопросы оптимальной упаковки $Cover(Cl)$ были рассмотрены в работе [6]). Другими словами каждой ненулевой точке расширенного образа $r_{ij}(Cl)>0$ соответствует черная точка $\widetilde{r}_{ij}(S) > 0$ в одном из образов $\widetilde{R}(S)$ символов, вошедших в кластер.

В предположении, что множество образов кластера Cl является репрезентативной выборкой для символа, породившего элементы кластера, определим понятие *идеального образа* как растра $R_{ideal}(Cl)$, на котором достигается минимум выражения:

$$\frac{1}{k}\sum_{p=1}^{k} \mu (R_{ideal} (Cl), R_p (S)) \rightarrow \min ,$$

где k – количество элементов в кластере Cl, а $S \in Cl$. В качестве функции близости μ может использоваться описанная выше симметрика d. Для дополнительных исследований кластера идеальный образ может быть распознан алгоритмом, обладающим точными и монотонными оценками, например нейронной сетью, описанной в [5].

Идеальный образ является одним из способов извлечения *эталона* символа – образа, применяемого в качестве элемента обучающей последовательности метода распознавании образов на втором проходе распознавания. Возможны иные способы извлечения эталона из множества образов кластера, оптимизирующие процессы самообучения и повторного распознавания с различными целевыми функциями, например, оптимизирующие быстродействие.

Основными характеристиками кластера $Cl =\{S_1, \ldots, S_k\}$ являются:

- расширенная область $Cover(Cl)$;

- идеальный образ $R_{ideal}(Cl)$;
- мощность кластера, то есть количество k составляющих его элементов.

Основные характеристики определены безотносительно характеристик символов S_j, поскольку описанный алгоритм цепной развертки использует образы, а не характеристики символов. Вообще говоря, алгоритм цепной развертки может построить кластеры на основе нераспознанных символов. Однако знание характеристик символа (начиная с кода символа и заканчивая кодом шрифта) существенно упрощает кластеризацию. Предположим, что для части образов символов перед началом кластеризации надежно определена некоторая характеристика, например, код символа. Мы оценивали надежность классификации кода символа по двум критериям:

- вероятность ошибки распознавания геометрического образа;
- подтверждение словарными механизмами распознанного достаточно длинного слова.

Используя в качестве функции расстояния между двумя образами x и y, с кодами символов S_x и S_y соответственно, симметрики

$$d_1(x,y) = \min(0, d(x,y) - \Delta\,\delta(S_x,S_y)),$$

где Δ – штраф за несовпадение кодов символов,

$\delta(x,y)$ – функция различия кодов, определяемая следующим образом:

$$\delta(x,y) = \begin{cases} 1, & \text{если} \quad x = y \\ 0, & \text{если} \quad x \neq y \end{cases},$$

и выбирая параметр Δ, мы добьемся совпадения кодов символов любого из образов, составляющих каждый из кластеров. При этом возможно появление кластеров малой мощности, состоящих из образов, классифицированных с ошибкой. Целесообразно *объединение* двух разноименных близлежащих кластеров с целью переименования одного из них. Оценка близости кластеров может производиться с помощью сравнения их расширенных областей метрикой в евклидовом пространстве, для переименования могут использоваться коды символов идеальных образов.

Аналогичными манипуляциями возможно построение кластеров таким образом, чтобы в одном кластере не было символов с различающимися характеристиками (признаки графем, код гарнитуры, код шрифта). То есть для любых двух символов S_A и S_B из одного кластера с характеристиками $\chi(S_A)$ и $\chi(S_B)$ справедливо выполнение одного из условий:

$\chi(S_A) = \chi(S_B)$,
или $\chi(S_A) = \theta$,
или $\chi(S_B) = \theta$.

Аналогично можно построить кластеры так, чтобы в каждом кластере содержались образы с одним и тем же значением кегля.

Поэтому имеют смысл определения следующих *дополнительных характеристик* кластера Cl, наследуемых из характеристик составляющих его символов:

- имя (код символов) кластера $C(Cl)$;
- кегль кластера $K(Cl)$;
- признаки графем $P_\lambda(Cl)$ (признаки жирности, серифности, наклона);
- код гарнитуры $G(Cl)$;
- код шрифта $F(Cl)$.

В каждом из кластеров Cl характеристики $P_\lambda(Cl)$, $G(Cl)$ и $F(Cl)$ могут быть неизвестны.

Определим отображение $\wp_1$, задающее принадлежность символа S и кластера Cl (элементов кластера $S_1,\ldots,S_n$) к одной группе однородности следующим образом.

$$\wp_1(S,Cl) = \begin{cases} 1, & \text{если} \quad \exists S_i \in Cl : \wp(S,S_i) = 1 \\ 0, & \text{в противоположном случае} \end{cases}$$

где $\wp$ - отношение принадлежности символов.

К дополнительным характеристикам кластера также относятся:

- оценка кластера, вычисляемая на основе оценок элементов этого кластера, например, как минимальная из оценок символов S_j, или как оценка распознавания идеального образа $R_{ideal}(Cl)$;
- надежность распознавания, вычисляемая на основе количества элементов кластера и количества символов кластера, подтвержденных словарным механизмом.

Представительными будем называть кластеры, подтвержденные по контексту, а также кластеры с достаточно большой мощностью или высокой оценкой.

Сомнительными будем называть кластеры, не являющиеся представительными.

Дадим определение шрифта, базирующееся на совокупности кластеров распознанных сим-

волов. *Шрифтом* ϕ мы будем называть такую совокупность кластеров символов, в которой для любых двух кластеров Cl_A и Cl_B, выполнены условия:

- коды символов кластеров различны;
- значения кегля кластеров одинаковы $K(Cl_A)=K(Cl_B)$;
- признаки графем кластеров одинаковы $P_\lambda(Cl_A)=P_\lambda(Cl_B)$

или $P_\lambda(Cl_A) = \theta$, но при этом
$\exists Cl_X \in \phi$: $\exists X \in Cl_X$, $P_\lambda(X) \neq \theta$, $\wp_1(S_A, X)=1$,

или $P_\lambda(Cl_B) = \theta$, но при этом $\exists Cl_X \in \phi$: $\exists X \in Cl_X$, $P_\lambda(X) \neq \theta$, $\wp_1(S_B, X)=1$;

- коды гарнитуры кластеров одинаковы $G(Cl_A)=G(Cl_B)$,

или $G(Cl_A) = \theta$, но при этом
$\exists Cl_X$: $\exists X \in Cl_X$, $G(X) \neq \theta$, $\wp_1(S_A, X)=1$,

или $G(Cl_B) = \theta$, но при этом
$\exists Cl_X$: $\exists X \in Cl_X$, $G(X) \neq \theta$, $\wp_1(S_B, X)=1$.

Отметим, что шрифты естественно представлять в виде мультимножеств, описанных в книге [7].

В нашем случае каждый шрифт ϕ представляет собой мультимножество $\{k_1 \bullet x_1, \ldots, k_s \bullet x_s\}$, носителем которого является алфавит распознавании $Supp(\phi)=\{x_1, \ldots, x_s\}$, а значение кратности k_i равно мощности кластера с кодом x_i.

Алгоритм поиска набора шрифтов на основе множества кластеров символов $Cl_1,\ldots,Cl_L$ аналогичен алгоритму, описанному в разделе 3 и базирующемуся на множестве символов $\overline{S}$.

Вообще говоря, кластеры символов $Cl_1, \ldots, Cl_L$ не могут быть разбиты на группы, соответствующие определению шрифта. Основная причина этого состоит в том, что использованный описанный выше метод кластеризации не гарантирует формирование единственного кластера для каждого символа алфавита шрифта, использованного при печати документа. Другой причиной являются ошибки и низкие оценки надежности распознавания характеристик символов, приводящие к появлению ошибок в дополнительных характеристиках кластеров.

Для целей последующего распознавания на основе построенных шрифтов, состоящих из кластеров символов, мы налагаем дополнительное требование: алгоритм построения шрифта должен максимизировать число представительных кластеров в шрифте.

Для построения шрифтов с максимизацией числа представительных кластеров можно использовать любой из методов кластеризации, например один из описанных в [2].

В качестве признаков, используемых для повторной кластеризации, мы предлагаем использовать дополнительные характеристики кластеров.

Важнейшим вопросом при кластеризации является выбор функции расстояния между объектами. В повторной кластеризации должна применяться функция, удовлетворяющая дополнительным требованиям. Учитывая, что в одном шрифте должен присутствовать только один кластер с определенным именем, расстояние между разными кластерами с одинаковыми именами должно быть велико. Поскольку, в первую очередь, хотелось бы отобрать для последующего использования на втором проходе представительные кластеры, то расстояние от представительного кластера до иного представительного должно быть меньше, чем до сомнительного кластера с остальными сходными параметрами.

Функция расстояния между кластерами Cl_A и Cl_B, удовлетворяющая перечисленным требованиям, в общем виде выглядит так

$$\rho_1(I_A, I_B) - Pen(Cl_A, Cl_B)$$

где I_A, I_B – идеальные образы кластеров Cl_A и Cl_B,

ρ_1 – функция сравнения идеальных образов, например, d,

Pen – функция штрафа за различие дополнительных характеристик кластеров.

При поиске шрифтов допустимо изменение набора кластеров символов, как вошедших в шрифты, так и не вошедших, с помощью следующих операций:

- переименование кластера, основанное на анализе шрифтов и распознавании идеального образа;
- уничтожение кластера, то есть вывод ненадежного кластера из рассмотрения;
- сумма двух разноименных близлежащих кластеров;
- разбиение кластера на два.

В двух последних операциях существенно представление кластеров в виде мультимножеств.

Поясним операцию разбиения на примере. Предположим, что в процессе неудачной кластеризации, в один кластер были ошибочно объединены символы, принадлежащим различным шрифтам (Рис. 2).

В результате анализа данного кластера он должен быть разбит на два отдельных кластера, приписанных различным шрифтам (Рис. 3).

5. Анализ кластеров с точки зрения алфавита

В случае отсутствия в построенном шрифте какого-либо кода кластера из алфавита распознавания требуется дополнительное исследование. Как хорошо известно, частота встречаемости разных букв существенно различна (Табл. 2, заимствованная из книг [8–9]).

Табл. 2.

Английский язык		Русский язык	
Символ	Частота (%)	Символ	Частота (%)
A	7.96	А	6.2
B	1.60	Б	1.4
C	2.84	В	3.8
D	4.01	Г	1.3
E	12.86	Д	2.5
F	2.62	Е,Ё	7.2
G	1.99	Ж	0.7
H	5.39	З	1.6
I	7.77	И	6.2
J	0.16	Й	1.0
K	0.41	К	2.8
L	3.51	Л	3.5
M	2.43	М	2.6
N	7.51	Н	5.3
O	6.62	О	9.0
P	1.81	П	2.3
Q	0.17	Р	4.0
R	6.83	С	4.5
S	6.62	Т	5.3
T	9.72	У	2.1
U	2.48	Ф	0.2
V	1.15	Х	0.9
W	1.80	Ц	0.4
X	0.17	Ч	1.2
Y	1.52	Ш	0.6
Z	0.05	Щ	0.3
		Ь, Ъ	1.4
		Э	0.3
		Ю	0.6
		Я	1.8

Отметим, что частоты существенно зависят не только от длины текста, но и от его характера. Например, в технических текстах редкая буква Ф может стать довольно частой в связи с частым использованием таких слов, как «функция», «дифференциал», «диффузия», «коэффициент» и т.п. Еще большие отклонения от нормы в частоте употребления отдельных букв наблюдаются в некоторых художественных произведениях, особенно в стихах.

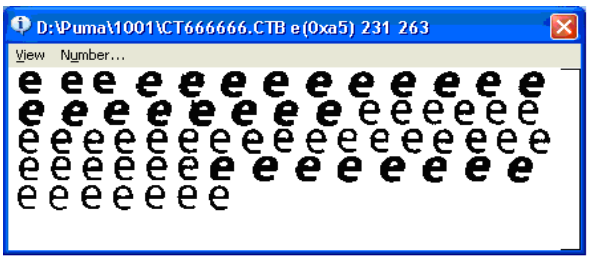
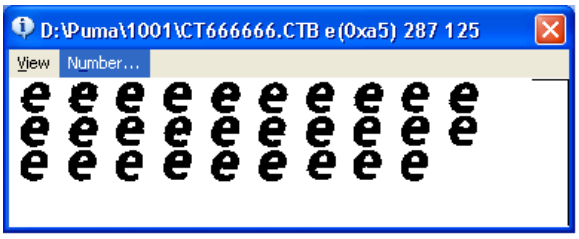


Рис. 2. Пример ошибочной кластеризации



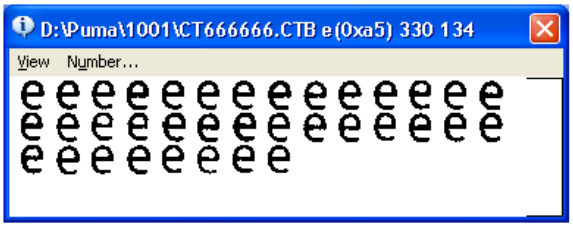


Рис. 3. Исправленные кластеры

Отсутствие среди отобранных кластеров для некоторого шрифта кластера с именем с высокой частотой встречаемости должно служить сигналом для пересмотра выбранных кластеров. Например, имя некоторого кластера могло быть определено неверно или кластер был ошибочно объединен с другим кластером с тем же именем (например, объединены кластеры жирного и светлого шрифтов).

Для дополнительной проверки построенных кластеров может быть использован метод, основанный на использовании критерия Романовского (определение см. в книге [10]) о соответствии теоретической и эмпирической функций распределения вероятностей.

Пусть $\alpha = \{\alpha_1, \ldots, \alpha_s\}$ – алфавит распознавания и $p=\{p_1,\ldots p_s\}$ – соответствующее распределение вероятностей появления символов алфавита, заданное с помощью частот встречаемости, приведенных выше в Табл. 2.

Рассмотрим шрифт $\bar{\phi} = \{n_1 \bullet \alpha_1, \ \ldots, \ n_s \bullet \alpha_s\}$. Пусть $n = n_1 + \ldots + n_s$.

Вычисляем величину $\rho = \dfrac{\left|\chi^2 - (s-1)\right|}{\sqrt{2(s-1)}}$, где $\chi^2 = \sum_{i=1}^{s} \dfrac{(n_i - np_i)^2}{np_i}$. В том случае, если $\rho \le 3$, то расхождение между эмпирическим и теоретическим распределениями считается несущественным (критерий Романовского).

Если величина $\rho > 3$, то расхождение между распределениями существенно. Определяем индекс $i^* = \{i : \max \dfrac{(n_i - np_i)^2}{np_i}, 1 \le i \le s\}$ и удаляем его из шрифта, полагая $k_{i^*} = 0$. Определяем для алфавита $\alpha^* = \{\alpha_1^*, \ldots, \alpha_{i^*-1}^*, \alpha_{i^*+1}^*, \ldots, \alpha_s^*\}$ новое распределение $p^* = \{p_1^*, \ldots, p_{i^*-1}^*, p_{i^*+1}^*, \ldots, p_s^*\}$ следующим образом:

$$p_j^* = \frac{p_j}{1 - p_{i^*}}, j = \overline{1,s}, j \ne i^*.$$

Полагаем $n^* = n - n_{i^*}$. Повторяем процедуру вычисления ρ для новых значений параметров, проверяем условие $\rho > 3$ и определяем очередной индекс, дающий максимальный вклад в расхождение. Такую процедуру можно повторить несколько раз с тем, чтобы найти в шрифте $\bar{\phi}$ кластеры, сомнительные с точки зрения частоты встречаемости букв.

Выявленные таким образом сомнительные кластеры должны быть обработаны одним из следующих способов:

- переименование в результате распознавания иным методом;
- разбиение на несколько кластеров;
- выведение из рассмотрения.

Отдельный анализ требуется для образованных маломощных шрифтов. Часто несколько слов в тексте бывает выделено особым образом – жирностью или курсивом. Если таких слов мало, то будет образовано несколько кластеров для данного шрифта, причем такие кластеры будут маломощные. Необходимо отличать такой случай от случайного объединения набора кластеров в ошибочный шрифт. Для этого можно применять усиленный анализ положения символов, входящих в кластеры. В случае действительного наличия шрифта слабой представительности положение символов в тексте должно быть хорошо согласовано – символы должны находиться в общих группах однородности.

6. Применение шрифтов

Построенный шрифт может быть использован для различных задач обработки документа. Мы кратко рассмотрим три задачи:

- повторное распознавание символов;
- упаковка изображения;
- подбор полиграфических шрифтов для показа результатов распознавания.

Первая из них имеет целью повторную сегментацию границ символов и повторное распознавание символов с низкими оценками и нераспознанных символов. Для сегментации на первом проходе, базирующейся на типовых отрезках разделения границ символов и шрифтонезависимом алгоритме распознавания образов символов (описание алгоритмов сегментации приведено в работе [11]), возможны следующие проблемы:

- ошибки, обусловленные неверными отрезками (кривыми) разделения символов;
- ошибки, связанные с артефактами шрифтонезависимого алгоритма распознавания.

Для устранения ошибок сегментации границ на втором проходе может применяться способ, использующий построенные кластеры и извлеченные из них эталоны.

Каждому из эталонов E соответствует образ $R(E)$, причем у каждого из эталонов размеры, то есть ширина $w(E)$ и высота $h(E)$, уникальны. Каждый из эталонов E обладает следующим набором характеристик $\chi(E)$:

- код символа $C(E)$, принадлежащий алфавиту распознавания A и являющийся кодом альтернативы распознавания с наибольшей оценкой надежности;
- кегль символа $K(E)$;
- признаки графем $P_\lambda(E)$ (признаки жирности, серифности, наклона);

- код шрифта $F(E)$ для эталонов, извлеченных из представительных кластеров.

Отметим, что для эталонов, приписанных к некоторому шрифту, ($F(E)\neq\theta$) определены все остальные характеристики $C(E)$, $K(E)$, $P_{\otimes}(E)$.

Объектом распознавания на втором проходе служит образ I, соответствующий последовательности из нескольких символов, которые были распознаны на первом проходе недостаточно надежно. Объект I является составной частью некоторой строки распознанных на первом этапе символов. Если часть символов распознанной строки классифицирована с точки зрения принадлежности построенным шрифтам, то на основе отношения $\wp$ образу I ставится в соответствие один или несколько шрифтов $\overline{F}(I)=\{F_1,\ldots,F_{n(I)}\}$. Возможен случай, когда $\overline{F}(I)=\varnothing$.

Сегментация образа I проводится следующим образов. Зафиксируем набор эталонов $\overline{E}=\{E_1,\ldots,E_Q\}$. В качестве $\overline{E}$ могут быть взяты следующие последовательности:

- все построенные эталоны;
- эталоны, принадлежащие всем построенным шрифтам;
- эталоны, принадлежащие нескольким построенным шрифтам;
- эталоны, принадлежащие одному шрифту.

Стратегия выбора набора эталонов $\overline{E}$ зависит от множества шрифтов-кандидатов $\overline{F}(I)$.

В соответствии с размером и правой границей каждого из эталонов E_i из распознаваемого образа I выделяется левая часть $L(I)$ так, чтобы размер $L(I)$ соответствовал размеру эталона E_i. После этого образ $L(I)$ сравнивается с эталоном E_i с помощью функции расстояния d. Процедура выделения части образа должна учитывать случаи невертикальных границ, например, в курсивном шрифте.

Таким образом, мы получаем некоторое количество вариантов образов левого начального символа $L_j(I)$, таких, что расстояние $d(L_j(I), E_i)$ меньше заданного порога. Для каждого варианта успешного распознавания левая часть удаляется из распознаваемого образа в соответствии с правой границей эталона, а с каждой из оставшихся частей I_j (для которой справедливо $I=L_j(I)\cup I_j$) операция повторяется. Процедура повторяется до тех пор, пока весь образ не будет распознан, или не будет получено, что приемлемых вариантов сегментации нет.

Вообще говоря, данная процедура является достаточно медленной. Механизм построения набора шрифтов на основе кластеризации, как средство уменьшения количества эталонов при сравнении, оптимизирует быстродействие.

Механизм построения набора шрифтов как в сегментации, так и в распознавании отдельного символа, повышает точность распознавания. Простым примером является повышение точности за счёт различения символов, обладающих сходными графемами в различных шрифтах (например, русская буква ‘З’ и цифра 3: ЗЗ 3З **3З)**и являющимися близкими по отношению к функции расстояния d.

Под *точность сегментации* будем понимать долю верно сегментированных слов к общему числу слов в тесовой последовательности. При этом под *верно сегментированными* словами понимаем слова, в которых правильно найдены границы всех символов, составляющих слово, вне зависимости от качества распознавания образов символов. Табл. 3 иллюстрирует повышение точности повторной сегментации за счет исправления ошибок алгоритмов сегментации на различных тестовых последовательностях (стендах) TS_1-TS_9.

Табл. 3. Точность сегментации после первого и второго прохода

	TS_1	TS_2	TS_3	TS_4	TS_5	TS_6	TS_7	TS_8	TS_9
без использования шрифтов	99,24%	99,83%	99,43%	99,76%	99,69%	99,40%	99,92%	99,32%	99,73%
с использованием шрифтов	99,84%	99,92%	99,87%	99,76%	99,76%	99,68%	100%	99,70%	99,86%

Задача упаковки изображения возникает как при сохранении результатов распознавания, так и при длительном хранении изображений. Предлагается формат хранения, состоящий из идеальных образов построенных шрифтов и описаний образов символов. Минимальное описание состоит из пары координат и ссылки на образ идеального символа. Рассмотрим пример простейшего образа отсканированного в разрешении 300 dpi документа, состоящего из 2000 символов одного единственного шрифта. Его компактное представление содержит образы идеальных символов (объемом менее 1 килобайта) и 2000 описаний координат и ссылок на шрифт (4 байта на описание). Без дополнительных ухищрений по упаковке мы достигли объема представления образа менее 10 килобайт. Для сравнения отметим, что объем черно-белого изображения A4 с разрешением 300 dpi, упакованного в формате TIFF Group 4, с аналогичными 2000 образов символов, составляет от 20 до 40 килобайт.

Задача выбора шрифтов является актуальной для каждой OCR, сохраняющей свои результаты в формате Microsoft Office. В идеале OCR должна уметь сохранять результат распознавания отсканированной страницы так, чтобы отображалась как структура страницы (разбиение на фрагменты текста, на текстовые строки, иллюстрации, таблицы), так и стили оформления распознанных символов. Последнее предполагает выбор шрифта для отображения группы распознанных символов, наиболее похожего на шрифт, использованный для печати данной группы символов. При это предполагается, что символы сгруппированы на основе сходства характеристик, таких кегль, признаки (наклон, жирность и т.п.), гарнитура, которые были определены выше в разделе 2.

Возможны две постановки задачи подбора шрифтов для отображения. В первой постановке необходимо выбрать из известного заранее множества шрифтов $\overline{F}_W$ шрифт с характеристиками, наиболее близкими по некоторой метрике к характеристикам группы однородных символов. Описанные выше алгоритмы распознавания шрифтов позволяют определить набор характеристик, по которым необходимо выбрать шрифт из множества $\overline{F}_W$. Разумеется, если $\overline{F}_W$ не содержит необходимо шрифт, то шрифт будет выбран приближенно с потерей ряда найденных характеристик шрифта. Например, полиграфическая гарнитура "Академическая" отсутствуют среди стандартных TTF-шрифтов Microsoft Windows, поэтому выбор возможен среди шрифтов типа Times New Roman, Tahoma и других серифных шрифтов. Часто применяется упрощенный подход, при котором множество шрифтов $\overline{F}_W$ содержит небольшое число различающихся друг от друга популярных гарнитур, таких как Times New Roman (серифный шрифт), Arial (бессерифный шрифт), Arial Narrow (узкие буквы), Courier (моноширинный шрифт). В упрощенном подходе характеристика гарнитуры игнорируется.

В другой постановке OCR сама строит шрифт в некотором формате (например, матричный шрифт), которым после регистрации этого шрифта в Microsoft Windows будет использован для отображения распознанных символов. Для этой постановке оказывается пригодным описанное распознавание шрифтов, дающее описание шрифта в виде множества идеальных образов.

Заключение

Использование в программах оптического распознавания текстов предложенных алгоритмов поиска шрифтов позволяет повышать качество и быстродействие алгоритмов адаптивного распознавания на основе интуитивно понятных моделей.

Автор выражает признательность В.Л. Арлазарову и А.Я. Подрабиновичу за поддержку исследований и плодотворное обсуждение.

Литература

1. Котович Н.В., Славин О.А. Алгоритмы распознавания шрифтов в печатных документах // В сб. трудов ИСА РАН "Обработка изображений и анализ данных Т.38. М.: Издательство ЛКИ, ", 2008, С. 252-271.
2. Арлазаров В.Л., Котович Н.В., Славин О.А. Адаптивное распознавание // Информационные технологии и вычислительные системы № 4, 2002, С. 11-22.
3. Арлазаров В.Л., Корольков Г.В., Славин О.А. Линейный критерий в задачах OCR // В сб. трудов ИСА РАН

"Развитие безбумажных технологий в организациях", М.: Эдиториал УРСС, 1999, С. 28-46.
4. Славин О.А. Распознавание атрибутов текстовых символов // сб. трудов ИСА РАН " Документооборот. Концепции и инструментарий", М.: Эдиториал УРСС, 2004, С. 142-150.
5. Гавриков М.Б., Мисюрев А.В., Пестрякова Н.В., Славин О.А. Об одном методе распознавания символов, основанном на полиномиальной регрессии // Автоматика и телемеханика. 2006, №3, С. 119-134.
6. Славин О.А., Титов Ю.В. Динамическое построение функций сравнения с идеальным образом в задаче адаптивного распознавания текстовых символов // Информационные технологии и вычислительные системы № 1, 2007, С. 3-12.
7. Петровский А.Б. Пространства множеств и мультимножеств. М.:УРСС, 2003.
8. Baudouin C. Elements de cryptographie. Ed. Pedone A. – Paris, 1939.
9. Яглом А. М., Яглом И. М. Вероятность и информация. М.: Наука, 1973.
10. Пиотровский Р.Г., Бектаев, К.Б., Пиотровская А.А. Математическая лингвистика. М.: Высшая школа, 1977.
11. Арлазаров В.Л., Куратов П.А., Логинов А.С., Славин О.А. Алгоритмы поиска границ печатных символов, используемые при оптическом распознавании символов, Информационные технологии и вычислительные системы № 4, 2004, С. 59-70.

Славин Олег Анатольевич. Заведующий лабораторией Института системного анализа Российской академии наук. Окончил МИРЭА в 1988 году. Кандидат технических наук с 2000 года. Автор 80 научных работ и изобретений. Область научных интересов: распознавание образов, искусственный интеллект. E-mail: oslavin@cs.isa.ru