УДК 004.91

А.А. Лигун, А.А. Шумейко, Д.В. Тимошенко

# О ЛОКАЛИЗАЦИИ И ФОРМИРОВАНИИ СИМВОЛОВ В ЭЛЕКТРОННЫХ ДОКУМЕНТАХ СО СЛОЖНЫМ ФОНОМ

## Введение

При современном развитии информационных технологий все большую значимость приобретает информация, получаемая посредством Интернет. Поэтому актуальным остается вопрос о сжатии, хранении и передаче электронных документов, анализе содержащейся текстовой информации, защите авторских прав, организации доступа и т.д. Существуют несколько форматов для хранения электронных растровых документов: DJVU[1], JPEG2000/Part 6[2] и LuraDocument (формат PDF является векторным, поэтому нами не рассматривается). В основу этих форматов положены идеи о разделении документа на цветовые плоскости, смысловую информацию и подложку. Заметим, что по своей сути эти форматы являются локальными, и их использование в сети сводится к передаче файла целиком. Фактически передается копия документа и дальнейшее ее использование проконтролировать сложно.

Авторами разработан подход к хранению и передаче электронных документов, который позволяет защитить информацию от несанкционированного копирования и снизить объемы передаваемых данных[3]. Данный подход реализован в виде формата ALD (ALLDocument), который ориентирован на использование в сети, но может быть использован и для работы на локальной машине. Метод сжатия документов ALLDocument основан на естественных принципах разделения на информативные слои. Предложенный подход, вместе с графическим форматом FWT, разработанным авторами, позволил получить формат, являющийся конкурентоспособным на рынке подобных продуктов.

В формате ALD заложена большая алгоритмическая гибкость, отражающаяся в замене и добавлении различных слоев информации и способов их кодирования. Каждый слой кодируется соответствующим обработчиком, который ориентирован на его

особенности. На данном этапе нами предусмотрена обработка следующих слоев:

1. Слой символьной информации.
2. Слой деловой графики.
3. Слой растровой информации.
4. Слой фона.

Представленные слои охватывают большинство типичных документов. Некоторые подходы к локализации слоя «деловая графика» нами рассматривались в работе [4].

Цель данной работы является описание первой части векторизации – локализации и формирования символов.

Существуют различные алгоритмы получения символьного слоя. При этом обычно предполагается, что символы расположены на однородном фоне, символы темнее – фон светлее, ярко выражены строки и столбцы, текст имеет горизонтальную ориентацию. Во многих случаях такие предположения отражают текущие характеристики обрабатываемого растрового документа. Однако круг документов, которые не охватываются подобным подходом, тоже весьма широк. Это различная полиграфическая продукция – иллюстрированные журналы, газеты, обложки книг и т.д.

Целью данной работы является разработка алгоритма локализации и формирования символьного слоя, который успешно сможет работать, как и в случае с однородным фоном, так и в случае сложного фона, в случае текста любой ориентации и любого соотношения его яркости и яркости фона.

Изначально нами предполагалось, что символ это часто повторяющееся связное множество точек со схожей яркостью и контрастное на окружающем фоне. Таким образом, формализация этих понятий является основой предлагаемого алгоритма.

Мы существенно применяли алгоритм, быстро приводящий к результатам, близким к оптимальному - квантованию. Для полноты изложения приведем некоторые факты, касающиеся оптимального квантования.

**Оптимальное квантование.** Квантование традиционно используется в алгоритмах сжатия числовых последовательностей с потерями. Суть квантования состоит в замене (в соответствии с заданными правилами) оригинальных значений функции

 ISSN 1562-9945

квантовочными числами. По ряду причин чаще всего в алгоритмах сжатия используется не оптимальное (квантование Ллойда-Макса [5, 6 с. 173]), а равномерное квантование. Вместе с тем, оптимальное квантование, являясь более трудоемким аппаратом приближения, дает наилучшее приближение данных кантовочными числами в среднеквадратичной метрике и с увеличением числа квантования ошибка квантования гарантировано не увеличивается.

Прежде чем остановиться на вопросе применения оптимального квантования при анализе числовых последовательностей, уделим внимание вопросу построения оптимальных квантовочных чисел. Известные нам алгоритмы, приводящие к квантованию Ллойда-Макса, как правило, обладают малой устойчивостью и большой ресурсоемкостью. Это как раз и является одной из причин отказа от оптимального квантования в пользу равномерного в алгоритмах сжатия.

В алгоритме локализации символов оптимальное квантование применяется в каждой точке изображения, поэтому вопрос скорости его работы и возможной оптимизации актуален. Авторами используется итерационный алгоритм, который путем уменьшения среднеквадратичной ошибки на каждом шаге, быстро дает результат, близкий к оптимальному. Что в итоге позволяет ограничиться заданным числом итераций квантования без критерия останова, зависящего от пересчитываемой ошибки квантования.

Пусть задана произвольная ограниченная функция дискретного аргумента $\varphi_i(i = 1,2,...,N)$ и $I_n$, который представляет собой вектор с координатами $b_{k+1/2}$ такими, что

$$\Phi^- = b_{1/2} < b_{3/2} < ... < b_{n+1/2} = \Phi^+,$$

где $\Phi^- = \min_i \varphi_i, \Phi^+ = \max_i \varphi_i.$

Положим $\Phi^+_{k+1/2} = \{\varphi_i \mid \varphi_i \le b_{k+1/2}\}$ $\Phi^-_{k+1/2} = \{\varphi_i \mid \varphi_i \ge b_{k+1/2}\}$.

Кроме того, пусть для ***k=1,2,..., n***

$$\Delta\Phi_k = \Phi^+_{k+1/2} \bigcap \Phi^-_{k-1/2}, \qquad (1)$$

Через $\delta_n$ обозначим множество фиксированных $\{b_k\}_{k=1}^n$ таких, что

$$b_1 < b_2 < .... < b_n, \text{ где } b_{k-1/2} < b_k < b_{k+1/2}.$$

Процедуру построения вектора $\delta_n$ будем называть правилом квантования функции $\varphi_i(i = 1,2,...,N)$ на n интервалов, числа $b_k$ -

квантовочными числами, $(b_{k-1/2}, b_{k+1/2})$- интервалами квантования. Замена функции $\varphi_i (i = 1,2,...,N)$кусочно-постоянной $\tilde{\varphi}$, такой, что $\tilde{\varphi}_i = b_k$ при $i : \varphi_i \in \Delta\Phi_k$, называется квантованием, при этом величина

$$\varepsilon(\varphi, \delta_n) = \sqrt{\frac{1}{N} \sum_{k=1}^{n} \sum_{i:\varphi_i \in \Delta\Phi_k} (\varphi_i - b_k)^2} \qquad (2)$$

называется ошибкой квантования.

Правило квантования, определяющее вектор $\delta_n^*$ такой, что

$$\varepsilon(\varphi, \delta_n^*) = \min_{\delta_n} \varepsilon(\varphi, \delta_n)$$

называется оптимальным квантованием или квантованием Ллойда-Макса.

Пусть $n_k$ - число элементов множества $\Delta\Phi_k$. При любых фиксированных интервалах квантования, оптимальные квантовочные числа $b_k$ для среднеквадратической метрики определяются равенствами

$$b_k = \frac{1}{n_k} \sum_{i:\varphi_i \in \Delta\Phi_k} \varphi_i. \qquad (3)$$

Заметим, что интервалы квантования, квантовочные числа, а также ошибка квантования не зависят от того, строятся они для функции $f$ или для ее неубывающей перестановки Харди. Поэтому, не теряя общности рассуждений, рассмотрим результаты для неубывающей функции.

**Теорема.** Пусть $f$ измеримая суммируемая в квадрате неубывающая функция. Обозначим через $m_{ab}(f) = \frac{1}{b-a} \int_a^b f(x)dx$ среднее значение функции на отрезке $[a,b]$. Тогда для любой точки $c \in [a,b]$ и такой $c^* \in [a,b]$, в которой выполняется условие $f(c^*) = \frac{m_{ac} + m_{cb}}{2}$, справедливо неравенство

$$\int_a^c (f(x) - m_{ac}(f))^2 dx + \int_c^b (f(x) - m_{cb}(f))^2 dx \geq$$

$$\int_a^{c^*} (f(x) - m_{ac^*}(f))^2 dx + \int_{c^*}^b (f(x) - m_{c^*b}(f))^2 dx .$$

Доказательство. Обозначим через

$$\Phi(c, m^-, m^+) = \int_a^c (f(x) - m^-)^2 dx + \int_c^b (f(x) - m^+)^2 dx .$$

 ISSN 1562-9945

Найдем такое значение $c^*$, при котором функция $\Phi(c, m^-, m^+)$ достигает своего минимума.

$$\frac{\partial \Phi(c, m^-, m^+)}{\partial c} = (f(c) - m^-)^2 - (f(c) - m^+)^2 = 0.$$

Таким образом, минимум достигается при $c^*$, для которого выполняется условие

$$f(c^*) = \frac{m^- + m^+}{2}.$$

То есть

$$\int_a^c (f(x) - m^-)^2 dx + \int_c^b (f(x) - m^+)^2 dx \geq \int_a^{c^*} (f(x) - m^-)^2 dx + \int_{c^*}^b (f(x) - m^+)^2 dx.$$

Найдем $m^{-*}, m^{+*}$, при которых $\Phi(c^*, m^-, m^+)$ достигает минимума.

$$\begin{cases} \dfrac{\partial \Phi(c^*, m^-, m^+)}{\partial m^-} = -2\left( \displaystyle\int_a^{c^*} f(x)dx - (c^* - a) \cdot m^- \right) = 0, \\ \dfrac{\partial \Phi(c^*, m^-, m^+)}{\partial m^+} = -2\left( \displaystyle\int_{c^*}^b f(x)dx - (b - c^*) \cdot m^+ \right) = 0. \end{cases}$$

Таким образом, минимум получаем при

$$m^{-*} = \frac{1}{c^* - a} \int_a^{c^*} f(x)dx, \; m^{+*} = \frac{1}{b - c^*} \int_{c^*}^b f(x)dx.$$

Следовательно,

$$\int_a^{c^*} (f(x) - m^-)^2 dx + \int_{c^*}^b (f(x) - m^+)^2 dx \geq \int_a^{c^*} (f(x) - m^{-*})^2 dx + \int_{c^*}^b (f(x) - m^{+*})^2 dx.$$

Теорема доказана.

Применяя процедуру, описанную в теореме, к произвольному методу квантования, на каждом шаге получаем новое квантование с гарантированно не увеличивающейся ошибкой (а в большинстве случаев – меньшей ошибкой).

Сформулируем итерационный алгоритм, приводящий к результатам, близким к оптимальному квантованию.

Для заданой ограниченной функции дискретного аргумента $\varphi_i (i = 1,2,..., N)$ первоначальный вектор $I_n$ зададим равномерным разбиением множества значений функции

$$b_{k+1/2} = \min_i \varphi_i + k \frac{\max_i \varphi_i - \min_i \varphi_i}{n}, \quad k = \overline{0, n}.$$

1. $k=1$.

2. Вычислим $\Delta\Phi_k$, $\Delta\Phi_{k+1}$ в соответствии с (1) и $b_k, b_{k+1}$ согласно (3).

3. Согласно теореме пересчитываем $(k+1/2)$-ую границу квантования

$$b_{k+1/2}=\frac{b_{k+1}+b_k}{2}.$$

3. $k=k+1$. Если $k<n$ – пункт 2.

4. Если ошибка квантования стабилизировалась – алгоритм останавливается, в противном случае – пункт 1.

Заметим, что в практических приложениях, где важна скорость работы алгоритма, можно обойтись без пересчета ошибки квантования после каждой полной итерации, а задаться наперед количеством полных итераций.

Локализация символов. Локализация символов заключается в выделении в отдельный слой пикселов, которые, скорее всего, принадлежат образу символа, и объединении полученных точек в множества, представляющие собой битовую маску символа (в дальнейшем просто символа). Локализацию символов будем проводить в два этапа.

*1 Этап.* Нахождение множества всех точек, принадлежащих символу или его окрестности (точки, полученные на этом этапе, будем называть точками символа первого рода). Формирование связных множеств $\{S_k^*\}_{k=1}^n$ из точек символов первого рода.

*2 Этап.* Построение на основе $\{S_k^*\}_{k=1}^n$ множеств $\{S_k\}_{k=1}^m, m\geq n$. Точки, принадлежащие множествам $\{S_k\}_{k=1}^m$, будем называть точками символа второго рода или просто точками символа.

Перейдем к более детальному описанию этапов локализации.

Будем работать с люминесцентной составляющей изображения (освещенность). Итак, пусть задана числовая матрица $Y=\{y_{i,j}\}_{i=1,j=1}^{W,H}$. При локализации символов исходим из основного предположения, что люминесцентная составляющая символа достаточно контрастна на окружающем фоне.

*1 Этап.* Будем считать точку изображения точкой символа первого рода, если:

Значения яркостей ее окрестности хорошо квантуются в два интервала при помощи оптимального квантования. То есть значения

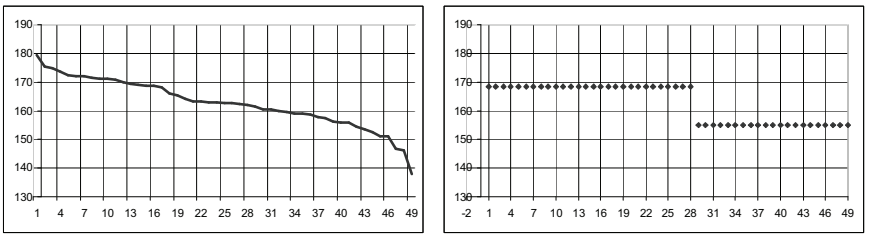
 ISSN 1562-9945

яркости окрестности группируются около ярко выраженных значений – средней яркости фона и символа.

Число точек, которые группируются около средней яркости фона, примерно равно числу точек, которые группируются около средней яркости символа – данная точка лежит на границе символа или около нее.

При этом, исходя из предположения о достаточной контрастности точек символа, полученные средняя яркость фона и символа существенно отличаются.

Пример поведения значений яркости окрестности точки символа первого рода приведен на рис 2, пример поведения точки, которая не является точкой символа первого рода – рис. 1.




а б

Рисунок 1 – а) типичный вид перестановки значений яркости в области точки, не являющейся точкой первого рода; б) типичный вид перестановки значений яркости в области точки, не являющейся точкой первого рода, отквантованная на 2 интервала (по оси X – номер точки в последовательности, по оси Y – ее яркость)

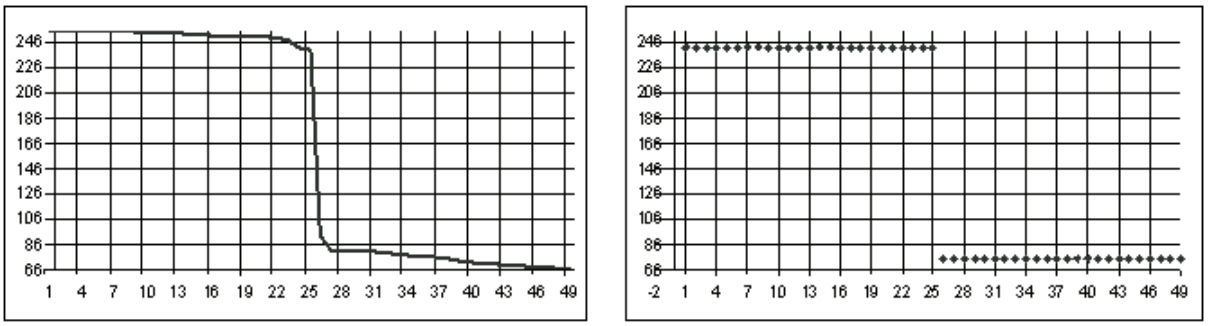



Рисунок 2 – а) перестановка люминесцентной составляющей окрестности точки символа первого рода; б) перестановка люминесцентной составляющей окрестности точки символа первого рода, отквантованная на 2 интервала (по оси X – номер точки в последовательности, по оси Y – ее яркость)

Для каждой точки с координатами $(i, j)$ определим окрестность $D_{i,j}^{N} = [i-N, i+N] \times [j-N, j+N]$. Для каждой фиксированной точки документа с координатами $(i, j)$ построим убывающую перестановку

яркости всех пикселов $y_{\nu,\mu}$ таких, что $(\nu,\mu)\in D_{i,j}^{N}$ (рис. 1а, рис. 2а). К последовательности яркостей применим оптимальное квантование на два интервала, результатом которого будет вектор $\delta_2=(b_1,b_2)$ Восстанавливая первоначальную информацию по квантовочным числам, получаем новую перестановку для заданной окрестности $\widetilde{y}_{\nu,\mu}$ (рис. 1б, рис. 2б).

Каждой точке $(i,j)$ поставим в соответствие число $\chi_{i,j}$

$$\chi_{i,j}=\frac{\sum\left\{1\,|\,(\nu,\mu)\in D_{i,j}^{N},\widetilde{y}_{\nu,\mu}=b_1\right\}}{(2N+1)^2}.$$

и относительную ошибку квантования

$$\varepsilon_{i,j}=\frac{\sqrt{\sum(y_{\nu,\mu}-\widetilde{y}_{\nu,\mu})^2\,|\,(\nu,\mu)\in D_{i,j}^{N}}}{\sqrt{\sum y^2{}_{\nu,\mu}\,|\,(\nu,\mu)\in D_{i,j}^{N}}}\cdot 100\%.$$

Зададим управляющие параметры $\delta_1>0,\delta_2>0,\delta_3>0,\delta_4>0$. Точка $(i,j)$ является точкой символа первого рода, если одновременно выполняются следующие условия

$$\varepsilon_{i,j}<\delta_1,$$

$$\frac{b_2-b_1}{b_2}\cdot 100\%>\delta_2,$$

$$\left|\chi_{i,j}\cdot 100\%-\delta_3\right|<\delta_4.$$

Оптимальные значения параметров $\delta_1,\delta_2,\delta_3,\delta_4$ подбираются путем имитационного моделирования или с помощью обучающих алгоритмов. Нами были выбраны такие значения параметров: $\delta_1=10\%,\delta_2=70\%,\delta_3=50\%,\delta_4=10\%$.

Завершением первого этапа локализации является построение связных множеств точек символа первого рода. Для этого нам понадобятся несколько определений.

Точки $(x_1,y_1)$ и $(x_2,y_2)$ будем называть смежными, если выполняется условие $|x_1-x_2|\le 1,|y_1-y_2|\le 1$.

Множество $S$ будем назвать связным, если $\forall(x_1,y_1)\in S$ и $(x_n,y_n)\in S$ $\exists\{(x_1,y_1),...,(x_i,y_i),(x_{i+1},y_{i+1}),....(x_n,y_n)\},(x_i,y_i)\in S$, где $(x_i,y_i),(x_{i+1},y_{i+1})$ - смежные точки.

 ISSN 1562-9945

Все множество точек символа первого рода разобьем на множество связных подмножеств при помощи одной из разновидностей алгоритма заливки[7].

Таким образом, итогом первого этапа локализации является получение связных множеств $\{S_k^*\}_{k=1}^n$.

*2 Этап*. Точки символа первого рода показали приблизительное расположение символа. Задача второго этапа по этим оценочным местам уточнить символ.

Заключим каждое множество $S_k^*$ в примитив $\mathrm{Pr}_k$. Например, таким примитивом может быть примитив, ограниченный внешней границей множества $S_k^*$. Заметим, что в процессе построения точек символа первого рода может произойти «склейка» слов в одно связное множество в виду малого размера символов в оригинальном тексте и как следствие малого расстояния между символами (рис. 3). Таким образом, в общем случае будем считать, что в примитиве могут находиться несколько символов.

а

б

Рисунок 3 – а) пример связного множества точек символа первого рода; б) пример примитива для связного множества точек символа первого рода

Считая, что точки символов, которые находятся в этом примитиве имеют схожую яркость, для получения яркости символов проводим оптимальное квантование на 2 интервала последовательности соотвествующих примитиву яркостей.

Для каждого примитива $\mathrm{Pr}_k$ построим убывающую перестановку яркости всех пикселов $y_{\nu,\mu}$ таких, что $(\nu,\mu)\in \mathrm{Pr}_k$. Применяя оптимальное квантование на 2 интервала, получим вектор $C_2=(c_1,c_2)$. Полученные квантовочные числа $c_1, c_2$ отражают среднюю яркость

фона и среднюю яркость символа. Однако изначально нам не известно отношение фона и символа в смысле яркости (символ темнее или светлее фона) и без дополнительного уточнения нельзя определенно сказать, что именно отображает каждое квантовочное число.

Для того чтобы определить какое из квантовочных чисел представляет собой яркость символа мы использовали информацию о точках, о которых уже достоверно известна их принадлежность фону. Такими точками являются точки, не попавшие ни в один примитив.

Для каждого примитива построим множество околограничных точек $Bg_k$, не принадлежащих самому примитиву, но имеющих в окрестности хотя бы одну точку из этого примитива. То есть

$$Bg_k = \{(i,j) \mid \forall l \quad (i,j) \notin \mathrm{Pr}_l \ , \ \exists (v,\mu) \in \mathrm{Pr}_k \quad (i,j) \in D^1_{v,\mu}\} \ .$$

Найдем среднюю яркость $Bg_k$

$$c_k^A = \frac{\sum y_{i,j} \mid (i,j) \in Bg_k}{\sum 1 \mid (i,j) \in Bg_k} \ .$$

Возвращаясь к кванновочному вектору яркостей примитива $\mathrm{Pr}_k \ C_2 = (c_1, c_2)$, считаем яркостью фона то значение, которое более другого отличается от средней яркости фона.

Пусть $|c_k^A - c_1| > |c_k^A - c_2|$. Тогда $c_1$ – средняя яркость символа в данном примитиве, и все точки, которые попали в соответствующий (в этом случае первый) интервал квантования считаются точками символа второго рода.

То есть точка $(i,j) \in \mathrm{Pr}_k$ считается точкой символа второго рода, если $\tilde{y}_{i,j} = c_1$ .

Все точки символа второго рода формируют связные множества $\{S_k\}_1^m$, что является финальным этапом приведенного алгоритма локализации и формирования символов. Заметим, что большинство символов, которые попали в одно связное множество на первом этапе, после второго разделяются на самостоятельные.

Пример текста с поэтапными демонстрациями локализации представлен на рис 4.

 ISSN 1562-9945


Рисунок 4 – а) оригинальный текст со сложным фоном; б) точки символа первого рода; в) точки символа второго рода

**Выводы**

Описан алгоритм локализации и формирования символов с применением оптимального среднеквадратичного квантования.

Предложенный алгоритм локализации и формирования символов был апробирован на реальных растровых документах со сложным фоном, произвольной ориентацией строк теста и различным яркостным соотношением символов и фона. Алгоритм показал себе как устойчивый к размерам символов, расположении, цвету.

Программная реализация алгоритма разработана таким образом, чтобы с подключением новых методов для обработки каждого слоя, техническая функциональность формата ALD не изменялась.

Этап локализации и формирования символов является первым этапом в обработке всего символьного слоя растровых документов. В данной статье мы не касались следующего этапа обработки символьного слоя – кластеризации и построения шаблона кластеров с последующим сжатием на основе словарных методов. Заметим, что вопрос формализации свойства символа многократного повторения в документе решается при кластеризации символов. В этом случае

символы, образующие кластер малой длины изымаются из данного слоя.

## ЛИТЕРАТУРА


1. Specification of DJVu Image Compression Format. – AT&T, 1999. – 39 p.
2. Information technology Jpeg2000 Image Coding System. Final Committee Draft .– 2006. – 205 p.
3. Лигун А.О., Шумейко О.О., Тимошенко Д.В. ALLDocument – технологія нового покоління для збереження, передачі та відображення електронних документів // Вісник Східноукраїнського національного університету імені Володимира Даля .– №9 (103) Частина 1 .– 2006 .– С. 83-85.
4. Лигун А.О., Шумейко О.О., Тимошенко Д.В. Локализация и формирование линий на изображении // Системные технологии. Региональный межвузовский сборник научных трудов. – № 3(50). – Днепропетровск. 2007. – С. 5-14.
5. Gray R., Neuhoff D. Quantization // IEE Transactions on Information Theory .– 1998 .– 44(6) .– P. 1-63.
6. Gersho A., Gray R. Vector quantization and signal compression. – Boston, 2001. – 730 p.
7. Роджерс Д., Адамс Дж. Математические основы машинной графики .– М.: Мир, 2001 .– 357 с.


Получено __.__.200_ г.

 ISSN 1562-9945