УДК 025.32:004.354

М. В. Дагаев

Применение формата DjVu в процессе оцифровки библиотечных фондов

Формат DjVu – один из способов сжатия изображений. В статье рассмотрены сильные и слабые места формата – их учёт позволяет более корректно планировать его использование для работы с результатами оцифровки.

Значительные объемы данных, получаемые при реализации проектов оцифровки печатных изданий и архивов, во многих случаях затрудняют хранение изображений «как есть», т.е. в несжатом виде. Поэтому постоянно разрабатываются способы сжатия изображений, позволяющие, с одной стороны, получить заметную экономию в объемах, а с другой – относительно невысокие потери качества в сжатых изображениях. Один из таких способов – формат DjVu, разработанный в AT&TResearchLabs.

Формат сжатия изображений DjVu (а также одноименная технология) основан на принципе MRC (*Mixed Raster Content* – растровое содержимое смешанного типа). В форматах сжатия графики, не использующих этот принцип (GIF, PNG, JPEG и т.д.), изображение сжимается как единое целое, вне зависимости от его характера. В форматах, использующих MRC, в том числе и в DjVu, исходное изображение перед собственно сжатием разделяется на несколько составляющих (далее – слоев) с содержимым одного и того же типа. После чего каждый из слоев обрабатывается теми алгоритмами сжатия, которые наиболее эффективны применительно к содержимому именно данного слоя.

Таких слоев обычно бывает два:

- “текстовая” часть изображения, для которой характерны резкие переходы между цветами (литеры текста, линии чертежа, штрихового рисунка и т.д.), обычно представлена в виде черно-белого изображения, но может быть и цветной (буквы плаката, цветные линии на карте, столбики диаграммы);
- полутоновая часть сжимаемого изображения с плавными, мягкими переходами цветов, без мелкой детализации (цветные иллюстрации, фон страницы).

Кроме самого разделения на слои, технология MRC позволяет хранить данные для каждого слоя с разными разрешениями. Это достаточно целесообразно, поскольку те части общего изображения, которые заключены в каждом из слоев, неравноценны по восприятию нашим зрением. Например, полутоновая часть – это обычно или фон страницы, или рисунки, не очень сложные по виду и цветовой гамме. Если их разрешение уменьшить в несколько раз по сравнению с исходным изображением, то результат, получаемый при выводе на экран, почти не будет отличаться от того, который мы имели бы, сохранив исходное разрешение. А вот объем той части DjVu-файла, в которой хранится информация об этом слое, сократится очень заметно.

Данные, находящиеся в “текстовой” части, обычно сохраняются с максимальным качеством, поскольку они должны не только обеспечить общую читаемость, но и сохранить точный вид литер текста, а также мелкие детали рисунков и чертежей.

В результате представленного подхода формат DjVu позволяет получать степень сжатия изображений в десятки раз большую, чем другие известные форматы сжатия – GIF, PNG, JPEG.

К сожалению, еще не создано способа разделения изображений, работающего с той же точностью, что и человеческое зрение, т.е. способного безошибочно отделять “текстовую” часть страниц от не-“текстовой”. Поэтому основная сложность, с которой приходится сталкиваться, применяя формат DjVu, – неаккуратное, т.е. неточное разделение общей структуры изображения на отдельные составляющие.

Само по себе неточное разделение на составляющие сложностей не создает. Но в дальнейшем у полутонового слоя разрешение будет уменьшено в несколько раз по сравнению с первоначальным. И если та часть изображения, которая по смыслу должна быть отнесена к текстовому слою, будет помещена в полутоновой, то вместо четких, хорошо различимых букв/линий она выведется на экран полуразмытым

пятном не всегда понятного содержания. Как такое выглядит, можно увидеть на примерах рисунков 1а и 1б. Рисунок 1а – это фрагмент иллюстрации из архитектурного альбома, рисунок 1б – то, что получилось после его преобразования в DjVu.

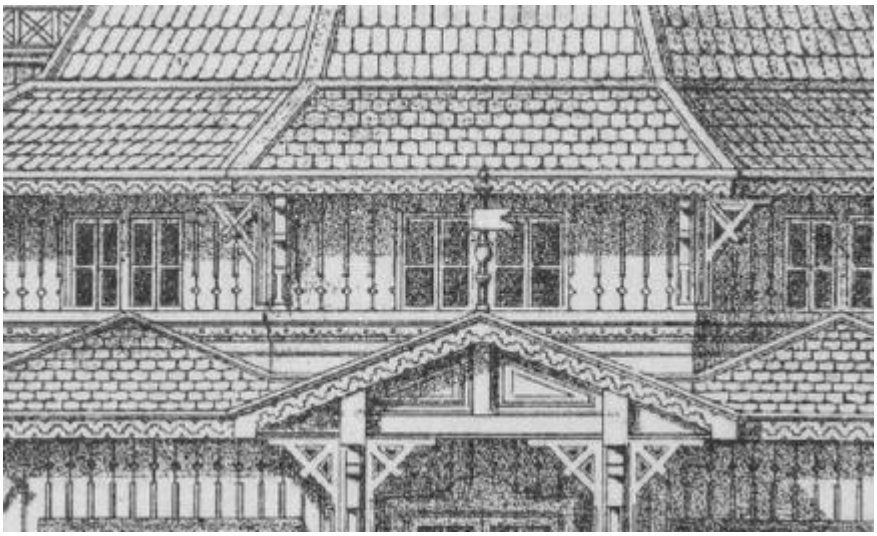

Рис. 1а. Фрагмент иллюстрации из альбома, посвященного архитектуре

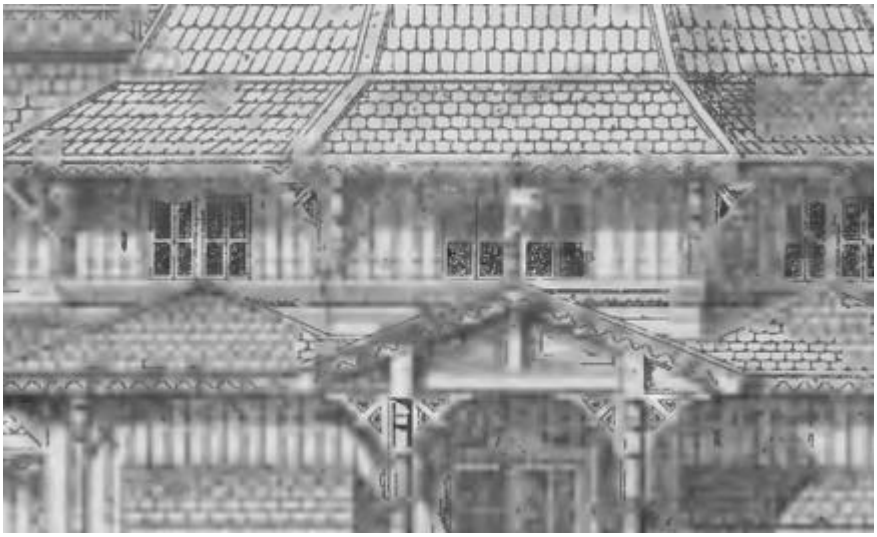

Рис. 1б. Фрагмент той же иллюстрации, преобразованный в DjVu

Основные требования к изображениям, сжимаемым в DjVu-формат

Изображения в цвете и в полутонах серого (Grayscale). Основное, что в этом случае необходимо DjVu-компрессору, – это возможность надежно отличить “текстовую” часть от не-“текстовой”, для чего желательно соблюдение двух основных условий:

1. Границы букв и линий рисунков должны быть четкими и неразмытыми.

Чем более резкая разница между фоном страницы и “текстовой” частью, тем больше вероятность того, что разделение на слои пройдет аккуратно и в результате получится DjVu-изображение, максимально соответствующее оригиналу.

2. Если изображение имеет вид “цветной текст на цветном фоне”, то необходимо, чтобы оттенки цветов в месте их соприкосновения были достаточно контрастными.

Вообще же одним из наилучших вариантов изображений для сжатия в DjVu можно считать текстовую страницу, отсканированную в цвете, но с фоном, по цвету близким к белому. В этом случае программа-компрессор практически безошибочно отделяет текст от фона, а фон переводит в чистый белый цвет. При этом можно получить степень сжатия, сравнительно с оригиналом, в районе 1:1000.

Нежелательно сжимать в DjVu:

а) изображения с множеством проведенных рядом тонких линий (сложнодеталированные штриховые рисунки, гравюры и т.д.).

Очень часто при разделении такие участки помещаются в полутоновой слой, разрешение в них понижается, из-за чего при сжатии утрачивается часть деталировки. Пример подобного некорректного преобразования – на рисунках 1а и 1б.

б) полутоновые изображения (фотографии, иллюстрации) с множеством мелких деталей.

В таких случаях очень часто основные контуры изображения остаются, но сложная деталировка при понижении разрешения размазывается и частично утрачивается.

Пример некорректной обработки фотоизображения – на рисунках 2а и 2б: рисунок 2а – это фрагмент фотографии из альбома видов старой Москвы, рисунок 2б – то, что из него получилось после преобразования в DjVu.

Рис. 2а. Фрагмент фотографии из альбома видов старой Москвы

Рис. 2б. Фрагмент той же фотографии после преобразования в DjVu

в) изображения размером более 450-500 Мб.

Начиная примерно с таких размеров, DjVu-программы очень часто отказываются сжимать файлы с изображениями. На попытку загрузить и сжать большое изображение они или зависают и перестают

реагировать на команды, или самопроизвольно прекращают работу и закрываются, т.е. реагируют тем или иным вариантом аварийного состояния.

Черно-белые изображения. Обработку черно-белых изображений можно считать одним из наиболее оптимальных вариантов использования формата DjVu: заложенные в нем алгоритмы позволяют сжимать черно-белые изображения в несколько раз более плотно, чем любые другие форматы сжатия.

В табл. 1 приведены объемы, которые занимает книга, сохраненная в виде черно-белых изображений, в разных форматах.

Таблица 1

Объем книги (с.)	Несжатый TIFF	TIFF Groupe 4	DjVu
256	378 Мб	19 Мб	3.8 Мб

Желательное условие для получения таких уровней сжатия – высокое качество сканирования. Разрешение сканирования должно составлять 400–600 dpi с правильно подобранной яркостью. Литеры текста на полученных страницах должны иметь ровно-аккуратный, легко читаемый вид, без разрывов или, наоборот, без мест, где линии букв сливаются в черные пятна.

Объёмы для книги, отсканированной с надлежащим качеством, даны в табл 1. Аналогичные данные для книги, отсканированной с не очень высоким качеством – в табл. 2.

Таблица 2

Объем книги (с.)	Несжатый TIFF	TIFFGroupe 4	DjVu
652	1690 Мб	113 Мб	46 Мб

Если в первом случае мы имели степень сжатия для TIFFGroupe 4 в 1:20, а для DjVu — 1:100, то во втором случае получим соответственно 1:15 и 1:38.

Программное обеспечение для работы с форматом DjVu

Для сжатия оцифрованных изображений в формат DjVu часто используют программу DjVu Solo. Ее основное достоинство в том, что это практически единственная некоммерческая версия DjVu-компрессора, выпущенная разработчиком программного обеспечения для данного формата – компанией *Lizardtech.*

Если требуется обрабатывать большие объемы изображений (сотни оцифрованных изданий за относительно короткие сроки), то наиболее оптимальным решением будет программа *Document Express Enterprise Edition 5.1.* Она создана как раз для задач поточно-конвейерной обработки больших объемов изданий. Но это коммерческая, т.е. платная программа.

Для просмотра полученных DjVu-файлов можно воспользоваться плагином к браузеру (если просматриваемый файл находится в Интернете), а также программой WinDjView (если просматриваемый файл находится на компьютере пользователя). И то и другое относится к некоммерческим программам.

Плагин, а также программы DjVu Solo и WinDjView легко доступны в Интернете.