ВЫЯВЛЕНИЕ JPEG-ИЗОБРАЖЕНИЙ СО ВСТРОЕННОЙ ИНФОРМАЦИЕЙ МЕТОДОМ СЖАТИЯ ФАЙЛОВ

А.И. Трубей, В.В. Кулага

Военная академия Республики Беларусь
220057, г. Минск-57, Республика Беларусь
телефон: + (37517) 287-41-29; факс: + (37517) 287-41-29; e-mail: trubeia@mail.ru
web: www.mod.mil.by

В статье описывается метод выявления встроенной информации в изображениях формата JPEG, основанный на особенностях сжатия файлов. С этой целью осуществляется вычисление коэффициентов сжатия исследуемых JPEG-файлов с помощью архиваторов и сравнение их с заданными пороговыми значениями. В зависимости от степени сжатия полученных архивных файлов делается вывод о наличии в исследуемом файле скрытой информации.

Ключевые слова – архиватор, стеганография, стегаграфический анализ, JPEG.

1 ТЕОРЕТИЧЕСКОЕ ОБОСНОВАНИЕ МЕТОДА

На IV международной конференции ITS'2008 был представлен доклад, в котором приводился метод выявления изображений формата JPEG со встроенными сообщениями, основанный на статистической оценке симметрии гистограмм коэффициентов ДКП [1]. В случае, когда гистограммы АС-коэффициентов ДКП изображений JPEG обладают достаточно высокой степенью симметрии, использование данного метода приведет к значительной ошибке первого рода. Поэтому можно воспользоваться более простым и удобным способом выявления таких файлов, который приводится ниже.

В своем труде "Математическая теория связи" Клод Шеннон смоделировал основы теории информации, в том числе идею о том, что данные могут быть минимизированы, так как несут избыточную информацию (энтропия данных) [2]. Энтропия исходных данных выступает количественной мерой разнообразия сообщений и является его основной характеристикой. Чем выше разнообразие алфавита сообщений и чем равномернее он распространен по тексту, тем больше энтропия и тем сложнее эту последовательность сообщений сжать.

Формат JPEG является самым распространённым и популярным стандартом компактного хранения полутоновых изображений (фотографий). Почти всегда используется так называемое сжатие с потерями, когда за счет небольшого ухудшения качества изображения значительно увеличивается степень сжатия и уменьшается размер конечного файла. Исходное изображение (BMP) сначала подвергается преобразованию ДКП. Потом выполняется квантование или округление. Потеря информации или незначительное ухудшение качества изображения происходит именно при квантовании. После этого данные сжимаются при помощи «энтропийного кодирования». В этом методе элементы данных, которые встречаются чаще, кодируются более коротким кодом, а более редкие элементы данных кодируются более длинным кодом. За счет того, что коротких кодов значительно больше, общий размер получается меньше исходного.

Поэтому использование архиваторов для повторного сжатия файлов в формате JPEG не дает значительного уменьшения размера, так как данные уже являются сжатыми. Тем не менее, даже в таких случаях сжатие теоретически возможно. Это обусловлено тем, что в большинстве распространенных форматов файлов, использующих сжатие, применены не самые эффективные методы. Например, в основе формата JPEG лежит «энтропийное» сжатие, в котором данные кодируются неоптимальными блоками, что обусловлено желанием сделать формат JPEG устойчивым к повреждениям и возможности частичного извлечения информации. Кроме самого изображения в файлах JPEG может храниться дополнительная информация, например, комментарии, параметры съёмки, а также параметры блока и файла, которые также обладают определенной избыточностью. Степень сжатия файлов характеризуется коэффициентом сжатия, определяемым как отношение объема сжатого файла V_c к объему исходного файла $V_и$: $K_c = V_c/V_и \times 100\%$. Чем меньше коэффициент сжатия, тем больше степень сжатия и эффективнее используемый метод сжатия.

Из экспериментальных данных известно, что в непрерывно-тоновых (многоградационных) изображениях сохраняется корреляция пикселов в соседних блоках размером 8×8. Пикселы таких изображений имеют величины, близкие значениям окрестных пикселов [3,4]. Следовательно, DC-коэффициенты $d_1,\ldots,d_n$, равные среднему значению коэффициентов блока, также являются коррелированными. Поэтому для естественных многоградационных изображений – фотографий, полученных съемкой объектов окружающего мира профессиональными фотоаппаратами с высоким разрешением и низким уровнем шумов матрицы, DC-коэффициенты соседних блоков будут не очень сильно различаться. При кодировании таких коэффициентов записывается первый (закодированный) DC-коэффициент, а затем кодируются разности DC-коэффициентов последовательных блоков [3,4]. Например, если первые три блока изображения имеют квантованные DC-коэффициенты 1118, 1114, 1119, то при JPEG-

сжатии записывается для первого блока число 1118 (закодированное кодом Хаффмана), за которым следует 63 закодированных АС-коэффициента. Для второго блока – число 1114 – 1118 = – 4 (закодированное тем же кодом) вперели 63 (кодированных) АС-коэффициентов блока. Третьему блоку будет соответствовать кодированная запись 1119 – 1114 = 5 и следующие 63 АС-коэффициента.

Таким образом, получим закодированную последовательность разностей DC-коэффициентов исходного изображения $d_{2,1}, d_{3,2}, \ldots, d_{i,i-1}, \ldots, d_{n,n-1} \subset D \subset Z$ ($d_{i,i-1}=d_i - d_{i-1}$), которая адекватно описывается стационарной цепью Маркова. Чем больше разрешение фотоаппарата, тем меньше число вариантов, принимаемых значениями разностей DC-коэффициентов последовательных блоков $d_{i,i-1}=d_i - d_{i-1}$ (мощность алфавита D), а, следовательно, меньше энтропия последовательности и выше частота повторения цепочек одинаковых байт.

При встраивании информации (случайной равновероятной последовательности) корреляция DC-коэффициентов соседних блоков будет нарушаться, поэтому последовательность разностей DC-коэффициентов будет иметь вероятностную модель, предусматривающую более равномерное их распределение. Число вариантов, принимаемых значениями разностей DC-коэффициентов (мощность алфавита D), будет большим. Следовательно, энтропия последовательности повысится.

Аналогичную корреляцию можно отметить также для АС-коэффициентов соседних блоков исходных изображений, которая также будет нарушаться при встраивании информации.

Энтропия любого отрезка информации равна вероятности его появления во всем массиве данных. Соответственно, наиболее часто повторяющиеся отрезки являются и наиболее «избыточными» и могут быть представлены в более сжатом виде [2]. Поэтому при архивировании изображений формата JPEG наиболее популярными архиваторами 7Zip, Zip, Rar коэффициент сжатия K_c для изображений со встроенной информацией будет больше, чем для исходных изображений.

Решающее правило для выявления факта наличия встроенной информации является следующим: осуществляется вычисление коэффициентов сжатия K_c исследуемых JPEG-файлов с помощью определенного архиватора и сравнение их с заданным пороговым значением K_0. Если $K_c \geq K_0$, то полагаем, что исследуемый файл содержит скрытую информацию и наоборот. Пороговое значение K_0 зависит от задаваемых параметров съемки и технических характеристик фотоаппарата. Для использования данного метода должны выбираться файлы, коэффициент сжатия которых не более 97-98%.

2 ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Осуществлялось вычисление коэффициентов сжатия архиватором 7Zip комплектов различных типов исходных фотографий и фотографий со встроенной информацией с применением стегосистемы «Steganos Security Suite – 2007», сделанных профессиональными фотоаппаратами с различными степенями разрешения, качества и сжатия. Использовались также изображения, полученные в результате обработки фотографий с применением растрового графического редактора «Photoshop». В качестве примера в таблице 1 приведены результаты сравнительного анализа коэффициентов сжатия. Метод дает результат с определенными вероятностями ошибок первого и второго рода: α – вероятность обнаружения скрытого сообщения в пустом контейнере, β – вероятность принятия контейнера с встроенным сообщением за пустой.

Таблица 1

Результаты архивирования комплекта файлов

Номер файла	Разрешение	K_c исходного файла (%)	K_c файла с встроенным сообщ. (%)
1	2126×2835	84	92
3	3189×2392	78	89
4	2126×3189	78	89
5	2126×3189	76	87
6	3189×2126	84	92
7	3189×2126	78	89
8	2362×3543	66	80
9	5315×3543	79	89
10	2362×3544	81	91
11	2950×2094	82	88
12	3508×2480	75	85
13	4966×3967	*86*	94
14	2592×3600	*73*	85
Суммарный комплект		80	90

Для большинства файлов, содержащих скрытую информацию, $K_c \geq K_0 = 85$. Факт наличия встроенной информации не подтвердился только для файла **8** (ошибка второго рода β). Кроме того, в пустом контейнере *13* ошибочно обнаружено наличие встроенного сообщения (ошибка первого рода α).

ЛИТЕРАТУРА

[1] Кулага В.В., Трубей А.И. Об одном методе статистической оценки симметрии гистограмм частот встречаемости коэффициентов ДКП // Информационные системы и технологии. Материалы IV Международной конференции (Минск, 4 – 8 ноября) – 2008. – С. 74–80.

[2] Shannon C. A mathematical theory of communication. - Bell system technical journal, 1948, v. 22, p. 379–423.

[3] Гонсалес Р., Вудс Р. Цифровая обработка изображений. – М.: Техносфера, 2005.

[4] Сэломон Д. Сжатие данных, изображений и звука.- М.: Техносфера, 2004.