Generalised Category Attack—Improving Histogram-Based Attack on JPEG LSB Embedding

Kwangsoo Lee[1], Andreas Westfeld[2], and Sangjin Lee[1]

[1] Center for Information Security Technologies (CIST),
Korea University, Seoul, Korea
kslee@cist.korea.ac.kr, sangjin@korea.ac.kr
[2] Technische Universität Dresden,
Institute for System Architecture,
01062 Dresden, Germany
westfeld@inf.tu-dresden.de

Abstract. We present a generalised and improved version of the category attack on LSB steganography in JPEG images with straddled embedding path. It detects more reliably low embedding rates and is also less disturbed by double compressed images. The proposed methods are evaluated on several thousand images. The results are compared to both recent blind and specific attacks for JPEG embedding. The proposed attack permits a more reliable detection, although it is based on first order statistics only. Its simple structure makes it very fast.

1 Introduction

LSB embedding is probably the most widespread and most frequently analysed steganographic method. It is instinctively considered weak. Hence many alternatives have been proposed in the last years. The JPEG file format is popular for digital photos not only in e-mail attachments. One of the first steganographic methods for JPEG files is Jsteg [1], which overwrites the LSB of DCT coefficients. The chi-square attack [2] demonstrated the weakness of Jsteg. Outguess [3] is an alternative that preserves the first order statistics and thus prevents the chi-square attack. It reserves a number of DCT coefficients to compensate for the changes. After embedding, the exact histogram is recovered by additional specific changes. This renders the chi-square attack ineffective, since it considers the histogram only. The chi-square attack is also prevented by F5 [4], which does not flip LSBs at all, but decreases the absolute value of coefficients by one, if necessary. F5 also increases the embedding efficiency, i. e., the message bits per change ratio. Another interesting approach is model-based steganography [5]. It uses an arithmetic decoder to adapt the stream of message bits to expected frequencies derived, e. g., from a Cauchy model of DCT coefficients. All the successors of Jsteg are immune against the chi-square attack. However, higher order statistical attacks, especially the complex blind attacks with a large feature space like the 324 Markov features by Shi et al. [6] and the 274 merged Markov and DCT

features by Pevný and Fridrich [7] can also detect the presence of embedded messages for the advanced steganographic JPEG methods.

Our motivation to consider again LSB embedding is twofold: With very low embedding rates and straddling the changes over the whole medium, LSB embedding is still hard to detect. This is not surprising at all. There are elaborate attacks for the spatial domain (Pairs analysis, RS, SPA [8,9,10]) and improved versions (scanning pixels along a space filling curve [11], adaptive hypotheses for natural images [12]). These are not easily applied to the DCT domain. It is unsatisfactory that the aforementioned complex blind attacks perform better than current specific attacks (Yu et al. [13], Zhang and Ping [14]). Recent measurements have shown that our previous approach [15] is not significantly better than the blind attack with 274 mixed features for images from one particular source. Jsteg length information can also be used to increase the detection power of targeted Jsteg attacks as shown by Westfeld [16], however, this improvement can easily be prevented by encryption. In practice, the images come from many different sources, are scaled and recompressed. This drops the performance of the blind attacks considerably. For our experiments, we use 900 images from CBIR [17], which is a very heterogeneous set of images and, since this set too small to train a classifier with more than 100 features, another set with 3000 never compressed TIFF images from NRCS [18]. We try to construct a similarly heterogeneous set from this second source, scale to different sizes and compress with different qualities. The aforementioned specific attacks do without a long training phase, without calibrated statistics, and without a large training set: They quickly return their result when directly applied to the suspect JPEG medium. In this paper we propose several improved versions of this attack, which seem to be less sensitive to double compression.

The paper construction is as follows: In the next section, we describe basic notations and definitions, and discuss AC JPEG distribution shapes. In Section 3, we propose a general version of the category attack proposed by Lee et al. [15]. In Section 4, the experimental results are presented for Jsteg and Jphide. Finally, we conclude this paper.

2 Preliminaries

2.1 Notations and Definitions

Let $h(x)$ denote the histogram of DCT coefficients in the cover image and let N be the total number of DCT coefficients. For the ease of description, we will use probabilistic terms. Let X be the random variable of DCT coefficients in a cover image, and $f(x)$ be the probability distribution of X, i. e., $f(x) = P_X(x)$. The relation between the two distributions is

$$f(x) = \frac{1}{N} h(x) \ . \tag{1}$$

Let X' be the random variable of DCT coefficients in a stego image in which a random message is embedded in the cover image with Jsteg. And let $f'(x)$ be the probability distribution of X' i. e., $f(x) = P_{X'}(x)$.

Let $U = \sum_{x \neq 0} h(x)$ denote the number of nonzero DCT coefficients (all AC and DC) and L the number of message bits to be embedded in the JPEG image. We will measure the capacity on a common basis: bits per nonzero coefficients or

$$\text{bpc} = \frac{L}{U} \ , \tag{2}$$

with $0 \leq \text{bpc} \leq 1$. The bpc is for all JPEG embedding algorithms the same while their particular capacities differ. For example, Jsteg is an application of the LSB embedding to JPEG format. It is identical with the LSB embedding except that the coefficient values 0 and 1 are not used for embedding. Consequently, for Jsteg, $U_1 = \sum_{x \notin \{0,1\}} h(x)$ and its relative capacity usage $\ell = \frac{L}{U_1}$.

We assume that the message bits are random. In other words, the message bits are assumed to be independent of the cover image and are uniformly distributed. Because the embedding path is pseudorandomly chosen, the probability of flipping the LSB of a coefficient $x \notin \{0, 1\}$ is $\ell/2$. Hence, one can establish a basic relation between the cover and the stego distributions as follows: for $x \neq 0$,

$$f'(2x) = f(2x) - \frac{\ell}{2} \left(f(2x) - f(2x+1) \right) \ , \tag{3}$$

$$f'(2x+1) = f(2x+1) + \frac{\ell}{2} \left(f(2x) - f(2x+1) \right) \ . \tag{4}$$

2.2 Consideration of AC JPEG Coefficient Distribution

The distribution of AC coefficients in JPEG images is often regarded to follow either the Laplacian distribution [19] or the generalised Cauchy distribution [5]. The Laplacian distribution

$$h(x) = \frac{\lambda}{2} e^{-\lambda |x|} \tag{5}$$

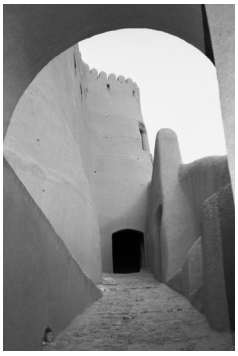
and the generalised Cauchy distribution

$$h(x) = \frac{p-1}{2s} \left(\left| \frac{x}{s} \right| + 1 \right)^{-p} \ . \tag{6}$$

have common properties: the symmetry about 0 and the unimodality.

The distributions AC coefficients of DCT transformed blocks are generally well approximated for images that are compressed for the first time. It is reasonable to believe that the quantisation does not affect the symmetry, however, we have found a considerable number of images with multimodal distribution, probably due to the effect of double compression. For example, Fig. 1 displays the quantisation table and the AC coefficient histogram of Y channel data of a JPEG image obtained from CBIR [17]. Clearly, the distribution is not unimodal. In CBIR, one can observe many JPEG images having multimodal "comb" distributions, which require special handling in steganalysis [20].

From this observation, the steganographic technique depends on the model parameters regarding the above distributions [5]. For example, the steganalytic



Quatisation table

	0	1	2	3	4	5	6	7
0	2	1	1	1	2	2	5	7
1	1	1	1	2	2	4	6	9
2	1	1	2	2	4	6	8	10
3	2	2	2	3	6	6	9	10
4	2	3	4	5	7	8	10	11
5	4	6	6	9	11	10	12	10
6	5	6	7	8	10	11	12	10
7	6	6	6	6	8	9	10	10

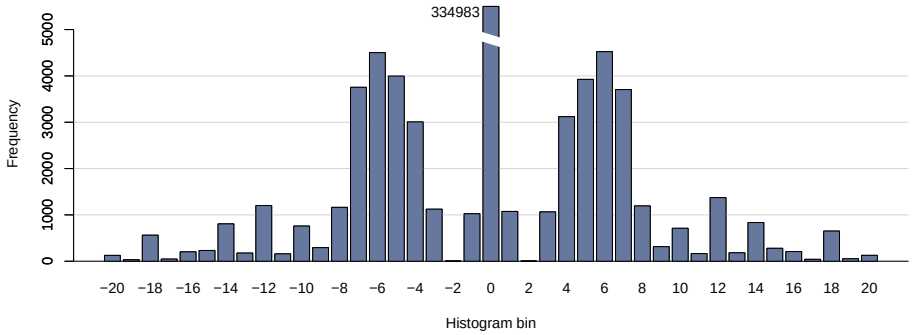


Fig. 1. A JPEG image from CBIR [17] and its quantisation table for Y channel and the histogram of AC coefficients showing the case of the multimodal; the frequency of the value 0 was reduced for display

technique by Yu et al. uses these model parameters [13]. One remedy for the steganalyst is to filter out the images of the unexpected case in the test domain. A more advanced one is to develop non-parametric solutions for the detection like the method of Zhang and Ping (ZP) [14] and the category attack by Lee et al. [15].

3 Generalised Category Attack

3.1 Histogram Modification for Detection

For the sake of simple description of the attack, we delete some bars in the histogram. We need two separate histogram modifications, one for Jsteg and another one for Jphide. After the modification, the histogram considers only values that are usable for LSB embedding.

Let $f(x)$ denote the probability distribution of the samples in a digital image. Then the relations of $f(x)$ and $f'(x)$ described in the Sect. 2.1 is true for all $x \in Z$.

We change the distribution in a way that keeps both, its smoothness as well as the effect of LSB embedding. Jsteg does not use the values 0 and 1. For a given histogram $h(x)$, we put

$$f(x) = \begin{cases} h(x)/U_1 & , \text{for } x > 1 \ , \\ h(x-2)/U_1 & , \text{for } x \leq 1 \ , \end{cases} \tag{7}$$

where U_1 is the number of usable coefficients, that is, $U_1 = \sum_{x \notin \{0,1\}} h(x)$. For the Jphide attack, we put

$$f(x) = \begin{cases} h(x)/U_2 & , \text{for } x > 1 \ , \\ h(x-3)/U_2 & , \text{for } x \leq 1 \ , \end{cases} \tag{8}$$

where $U_2 = \sum_{x \notin \{-1,0,1\}} h(x)$ ignoring occasional changes and the negligible cases of usable coefficients.

3.2 The Concept of the Category Attack

Consider two pairs of values: $(2m, 2m+1)$ and $(2m-1, 2m)$. The former is named the *induced* category in a sense that LSB embedding induces the decrement of the frequency difference,

$$|f'(2m) - f'(2m+1)| = (1-\ell)|f(2m) - f(2m+1)| \ , \tag{9}$$

which is derived from (3) and (4). This fact was earlier used in the chi-square attack [2,21], a categorical data analysis for the detection of steganography.

The latter is named the *shifted* category in a sense that the values are shifted by 1. Induced and shifted categories have been termed direct and shifted pairs in the literature before [8]. However, apart from the correspondence of these terms, the category attack has little resemblence with the Pairs analysis that exploits higher order statistics creating two binary vectors and evaluating their number homogeneous and inhomogeneous pairs. The main idea of the category attack is to use the shifted category in which the frequency differences have a different pattern with LSB embedding compared to that in the induced category: Because the values can be changed within an induced category containing it, the effect of LSB embedding on the frequency difference is covered by at least two induced categories, $(2m-2, 2m-1)$ and $(2m, 2m+1)$, which are overlapped by the shifted category $(2m-1, 2m)$. In other words, the effect is dependent on the state of the distribution $f(x)$ on the four consecutive sample values $\{2m-2, 2m-1, 2m, 2m+1\}$.

The assumed sample distribution is varying, and thus, we expect that $f(x)$ is monotonically increasing or decreasing on many intervals with a significant portion. For the JPEG domain, this is clear regardless of what the modality of the distribution is. Consider a monotone decreasing interval. The equalisation of both frequencies in the induced category implies that $f(2m-1)$ will increase and $f(2m)$ will decrease after LSB embedding. This means that the frequency

difference in the shifted category $(2m-1, 2m)$ grows when the embedded message size is increasing. This is also true for the monotone increasing case. This makes a difference to the pattern in the induced category.

In Fig. 2, the left side shows the difference between both changes of the frequency difference in the induced and the shifted categories. The curves represent the monotone decreasing function $f(x)$. Two adjacent circles with the same colour represent the values in an induced category. For each circle, the arrow indicates the changing pattern of the frequency of the sample value by LSB embedding. The top-left and the bottom-left figure show both patterns in the induced and the shifted categories respectively. The two arrows in one category point in opposite directions. The category attack evaluates these differences in a relative way using some measurements.

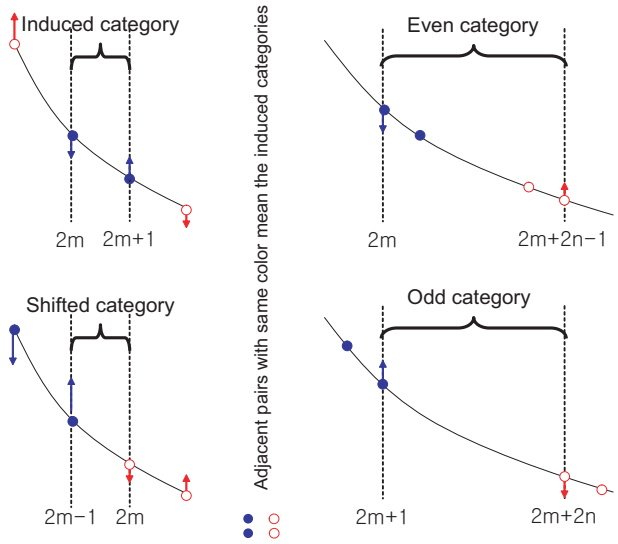


Fig. 2. The concept of general categories

3.3 General Categories

In this section, we will generalise the concept of the category attack. A category means a pair of values, say (a, b), such that $a, b \in Z$ and $a < b$. It is more convenient to represent the category by the small value a and the distance $d = b - a$. Let Ω_d be the collection of categories in which values differ by d, i. e., $\Omega_d = \{(a, a+d) : a \in Z\}$. For a distance d, the categories in Ω_d can be classified into LSB types of their small values. Let $\Omega_{0,d}$ be the sub-collection of Ω_d whose members are of the form $(2m, 2m + d)$, $m \in Z$, i. e., categories having even values as their small values. Similarly, let $\Omega_{1,d}$ be the sub-collection of Ω_d whose

members are of the form $(2m+1, 2m+d+1)$, $m \in Z$, i. e., categories having odd values as their small values. Then the collection Ω_d is the disjoint union of the sub-collections, $\Omega_{0,d}$ and $\Omega_{1,d}$.

When the distance d is an odd value, $\Omega_{0,d}$ and $\Omega_{1,d}$ have some distinctive characteristics under LSB embedding. In this case, say $d = 2n-1$ for some $n \in N$, the category collections can be rewritten as $\Omega_{0,2n-1} = \{(2m, 2m+2n-1) : m \in Z\}$ and $\Omega_{1,2n-1} = \{(2m+1, 2m+2n) : m \in Z\}$. We will refer to $\Omega_{0,2n-1}$ as the collection of even categories with a distance $2n-1$ (briefly the even categories) and to $\Omega_{1,2n-1}$ the collection of odd categories with a distance $2n-1$ (briefly the odd categories) respectively. The induced categories and the shifted categories referred to in the previous section are the special case of the even and the odd categories with the distance 1, respectively.

Let us look again at Fig. 2. The right side displays the different changing patterns (arrows) of the frequency differences in the even and the odd categories by LSB embedding. For the even category $(2m, 2m+2n-1)$ with a distance greater than 1, the two induced categories overlapped by the even category are $(2m, 2m+1)$ and $(2m+2n-2, 2m+2n-1)$. Consider the monotone decreasing interval with $f(x)$ again. As shown in the top-right figure, the frequency $f(2m)$ will decrease and the frequency $f(2m+2n-1)$ will increase after LSB embedding. Their difference $f(2m) - f(2m+2n-1)$ will decrease. On the other hand, for the odd category $(2m+1, 2m+2n)$ with a distance greater than 1, both induced categories overlapped by the even category are $(2m, 2m+1)$ and $(2m+2n, 2m+2n+1)$. Considering a monotone decreasing interval of $f(x)$ again (cf. bottom-right in the figure), the frequency $f(2m+1)$ will increase and the frequency $f(2m+2n)$ will decrease after LSB embedding. Their frequency difference $f(2m+1) - f(2m+2n)$ will increase. The frequency difference in the even and the odd categories will grow with the LSB embedding rate. And this is also true for a monotone increasing interval. Therefore, the even and the odd categories can be viewed as the generalised concept of the induced and the shifted categories.

3.4 Measurements of the Frequency Differences in Categories

For each type of categories with a distance $2n-1$, an easy calculation of the frequency difference as overall statistics is the sum of absolute values of the frequency differences in the categories:

$$S_1(0,n) = \sum_{x \in Z} |f(2x) - f(2x+2n-1)| \,, \tag{10}$$

$$S_1(1,n) = \sum_{x \in Z} |f(2x+1) - f(2x+2n)| \,. \tag{11}$$

These have linear patterns of the embedding rate ℓ in their change. Another measure is of the form of the squared sum of frequency differences:

$$S_2(0,n) = \sum_{x \in Z} (f(2x) - f(2x+2n-1))^2 \,, \tag{12}$$

$$S_2(1,n) = \sum_{x \in Z} (f(2x+1) - f(2x+2n))^2 \ . \tag{13}$$

In general, the p-th powered sum of the absolute values of the frequency differences can be considered. However, our experiments have shown no improved detection power compared with the above measures. The following χ^2-like measure was used in the original version of the category attack:

$$\chi^2(0,n) = \sum_{x \in Z} \frac{(f(2x) - f(2x+2n-1))^2}{f(2x) + f(2x+2n-1)}, \tag{14}$$

$$\chi^2(1,n) = \sum_{x \in Z} \frac{(f(2x+1) - f(2x+2n))^2}{f(2x+1) + f(2x+2n)}. \tag{15}$$

After LSB embedding, we would expect that

$$S_i'(0,n) < S_i(0,n) \ \text{ but } \ S_i'(1,n) > S_i(1,n), \tag{16}$$

$$\chi^{2'}(0,n) < \chi^2(0,n) \ \text{ but } \ \chi^{2'}(1,n) > \chi^2(1,n), \tag{17}$$

where $i \in \{1,2\}$. Furthermore, we expect each measurement to be more sensitive to higher embedding rates ℓ; for example, the change of $S_1(0,n)$ and $S_1(1,n)$ is linearly decreased and increased on ℓ respectively, if $f(x)$ is increasing in all intervals.

For a fixed distance $2n-1$, we compare both measurements in the even and the odd categories in a relative way. Put

$$\mathrm{R}_{S_i}(n) = \frac{S_i(1,n) - S_i(0,n)}{S_i(1,n) + S_i(0,n)} \ , \tag{18}$$

$$\mathrm{R}_{\chi^2}(n) = \frac{\chi^2(1,n) - \chi^2(0,n)}{\chi^2(1,n) + \chi^2(0,n)} \ , \tag{19}$$

We assume that cover images have a stable pattern of the relativities, $\mathrm{R}_{S_i}(n)$ and $\mathrm{R}_{\chi^2}(n)$. If the relativity is above a threshold, then we suppose a stego image.

Let us consider the distance $2n-1$. We have also evaluated the following combined relativities:

$$\mathrm{CR}_{S_i}(n) = \frac{\sum_{k=1}^{n} S_i(1,k) - \sum_{k=1}^{n} S_i(0,k)}{\sum_{k=1}^{n} S_i(1,k) + \sum_{k=1}^{n} S_i(0,k)} \ , \tag{20}$$

$$\mathrm{CR}_{\chi^2}(n) = \frac{\sum_{k=1}^{n} \chi^2(1,k) - \sum_{k=1}^{n} \chi^2(0,k)}{\sum_{k=1}^{n} \chi^2(1,k) + \sum_{k=1}^{n} \chi^2(0,k)} \ . \tag{21}$$

When n is growing, it is reasonable that the relativity becomes more stable for cover images. However, at the same time the relativity becomes less sensitive to LSB embedding as our experiments have shown. There seems to exist some trade-off between the stability and the sensitivity.

3.5 Discussion of the Applications to Jsteg and Jphide

Assume that JPEG images have the symmetry around 0. (7) yields

$$f(1+x) = h(1+x)/U_1 = h(-1-x)/U_1 = f(1-x) \tag{22}$$

This means that the symmetry is also true for the modified distribution with Jsteg but the center is changed to 1. Let $y = 2 - 2x - 2n$, then $|f(2x) - f(2x + 2n - 1)| = |f(2y + 1) - f(2y + 2n)|$, and thus,

$$S_i(0, n) = S_i(1, n) \ . \tag{23}$$

Similarly, one can deduce

$$\chi^2(0, n) = \chi^2(1, n) \ . \tag{24}$$

We have

$$\mathrm{R}_{S_i}(n) = 0 \ \text{ and } \ \mathrm{CR}_{S_i}(n) = 0 \ . \tag{25}$$

$$\mathrm{R}_{\chi^2}(n) = 0 \ \text{ and } \ \mathrm{CR}_{\chi^2}(n) = 0 \tag{26}$$

As an improvement of the category attack for Jsteg, instead of $h(-1)$, we suggest to use $h(1)$ for the calculation in the odd categories. So we introduce a new histogram modification $f_{\mathrm{ICA}}(x)$ that is equivalent to those defined in Eqns. (7) and (8) for the cover distributions except for $f_{\mathrm{ICA}}(1) = h(1)/U_1$. Due to the symmetry $h(x) = h(-x)$ this will not change the cover statistics. However, because $h(1)$ is not changed while $h(-1)$ decreases after embedding, the measurement of frequency difference in the odd categories will be greater than its original version after LSB embedding. This will boost the sensitivity to Jsteg. We will call this improved category attack (ICA).

If the cover distribution has no symmetry around a sample value, the relativity can be hardly predictable. The initial quantities of the relative differences will fluctuate much more and consequently deviate from the mean. This is the reason of why the category attack is worse for Jphide than Jsteg. Using the combined version of the generalised category attack (GCA), this will be more improved. However, there is no better way of guessing the parameter n than by experiments.

4 Experimental Results

The results are based on two image sources. About 900 JPEG images were downloaded from the CBIR [17]. These 900 images are from mixed sources, different size, colour and greyscale, and possibly double compressed. These images were classified by all versions of the specific attacks.

For the blind attacks we faced the same problem that an investigating officer has if he only has a small set of images to test but no sufficiently large training set. A set of 900 images is too small to be separated into subsets for training and test

Table 1. Selected reliabilities and false positive rates of the proposed GCA and ICA methods in comparison to existing specific and blind attacks

		bpc=0.01		bpc=0.02		bpc=0.04	
Attack	$2n-1$	ρ	$\text{FPR}_{0.5}$	ρ	$\text{FPR}_{0.5}$	ρ	$\text{FPR}_{0.5}$
Jsteg						(900 CBIR images)	
GCA_{AC} CR_{χ^2}	1	0.334	0.239	0.599	0.094	0.877	0.012
GCA_{AC} CR_{χ^2}	3	0.366	0.240	0.643	0.078	0.911	0.010
GCA_{AC} CR_{χ^2}	5	0.341	0.237	0.612	0.098	0.897	0.020
GCA_{AC} CR_{χ^2}	7	0.313	0.280	0.573	0.130	0.876	0.027
GCA_{AC} CR_{χ^2}	9	0.288	0.305	0.535	0.152	0.846	0.033
ICA_{AC} CR_{χ^2}	1	0.353	0.265	0.632	0.127	0.898	0.028
ICA_{AC} CR_{χ^2}	3	0.277	0.319	0.513	0.191	0.805	0.067
ICA_{AC} CR_{χ^2}	5	0.250	0.323	0.471	0.220	0.759	0.084
ICA_{AC} CR_{χ^2}	7	0.239	0.331	0.452	0.228	0.738	0.099
ICA_{AC} CR_{χ^2}	9	0.230	0.332	0.437	0.234	0.720	0.109
CA (GCA_{DC} CR_{χ^2}	1)	0.322	0.251	0.583	0.098	0.866	0.012
ZP β		0.193	0.343	0.366	0.252	0.632	0.140
ZP_{AC} β		0.198	0.349	0.376	0.249	0.645	0.138
Yu α		0.056	0.487	0.110	0.477	0.209	0.441
Yu α w/o zero bin		0.059	0.467	0.118	0.444	0.228	0.394
23 DCT (NRCS)		0.062	0.455	0.129	0.411	0.247	0.301
324 Markov (NRCS)		0.075	0.456	0.151	0.407	0.271	0.339
274 Merged (NRCS)		0.234	0.315	0.432	0.165	0.687	0.016
Jphide						(900 CBIR images)	
GCA_{AC} CR_{S_2}	1	0.154	0.361	0.296	0.250	0.526	0.119
GCA_{AC} CR_{S_2}	3	0.163	0.404	0.313	0.286	0.564	0.112
GCA_{AC} CR_{S_2}	5	0.178	0.388	0.344	0.272	0.623	0.095
GCA_{AC} CR_{S_2}	7	0.166	0.408	0.318	0.301	0.577	0.108
GCA_{AC} CR_{S_2}	9	0.163	0.406	0.313	0.330	0.570	0.118
ICA_{AC} CR_{χ^2}	1	0.172	0.343	0.322	0.222	0.556	0.115
ICA_{AC} CR_{χ^2}	3	0.176	0.430	0.327	0.325	0.571	0.128
ICA_{AC} CR_{χ^2}	5	0.172	0.443	0.316	0.363	0.545	0.185
ICA_{AC} CR_{χ^2}	7	0.163	0.448	0.303	0.378	0.525	0.201
ICA_{AC} CR_{χ^2}	9	0.156	0.430	0.294	0.370	0.516	0.212
CA (GCA_{DC} CR_{χ^2}	1)	0.151	0.383	0.293	0.267	0.536	0.157
23 DCT (NRCS)		0.005	0.493	0.015	0.488	0.034	0.478
324 Markov (NRCS)		0.136	0.420	0.258	0.354	0.386	0.272
274 Merged (NRCS)		0.159	0.378	0.305	0.283	0.524	0.182

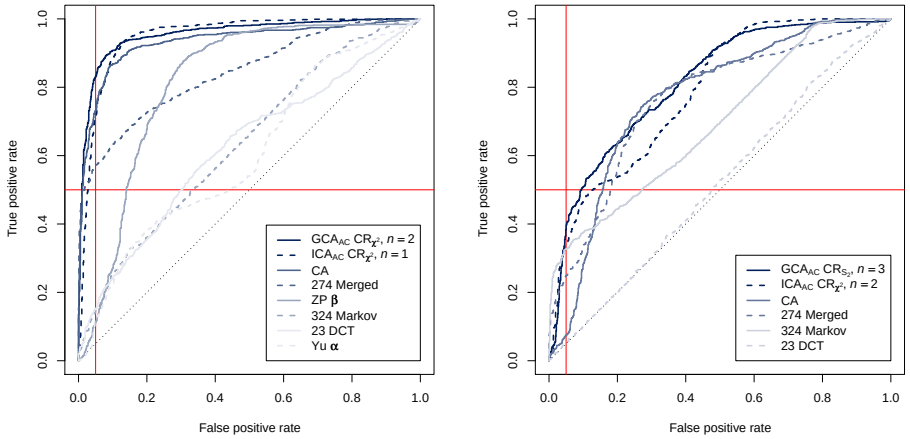


Fig. 3. ROC curves based on 900 CBIR images [17] for selected attacks to 0.04 bpc Jsteg (left) and 0.04 bpc Jphide (right). The three blind classifiers have been trained on a reasonable subset of 100,000 images derived from NRCS images [18]

with about 300 features. So another 3000 large colour TIFF images (2100×1500) were downloaded from the NRCS Photo Gallery [18], to construct a training set. We downscaled the images using pnmscale (smaller side 600, 400, 200, 80, and 40 pixels), and converted to greyscale JPEG using pnmtojpeg (qualities 99, 95, 90, 80, 70, 60, 50). Sufficiently large subsets of these about 100,000 images[3] were used to train the blind classifiers. We implemented the blind attacks by Fridrich with 23 DCT features [22], by Shi et al. with 324 Markov features [6], and by Pevný and Fridrich with 274 merged extended DCT and reduced Markov features [7], the specific attacks on randomised Jsteg by Yu et al. (the model-based approach) [13], the attack by Zhang and Ping [14], and by Lee et al. [15], together with their attack on Jphide. We focussed on low embedding rates (0.01, 0.02, and 0.04 bits per nonzero coefficient [bpc]).

Table 1 shows selected results for the attacks on Jsteg and Jphide. Figure 3 presents the ROC curves for the proposed attacks (GCA and ICA) with their best selection of n together with previous attacks. The ROC curve of some attacks (e. g., the attack by Yu et al. based on their α value) is very sensitive to the heterogeneous composition of the image set. These curves show cavities that disappear if the same attack is applied to a homogeneous set with equally sized images from one source. The detection power is evaluated using two measures: one is the reliability ρ of the ROC curve, which is twice the area under the curve minus one, and $\mathrm{FPR}_{0.5}$, the false positive rate (FPR) at true positive rate (TPR) 0.5. We applied the three blind attacks (23 DCT, 324 Markov, and 274 merged

[3] The size of the training set was chosen according to the number of features, not smaller than 10 images per feature and class.

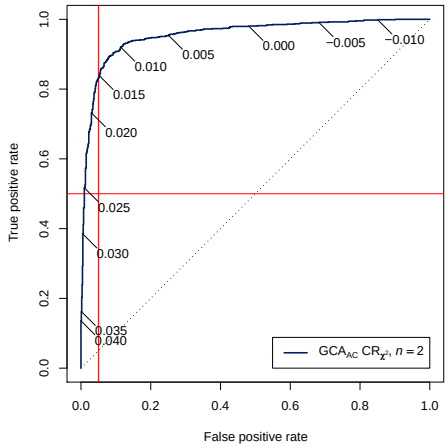


Fig. 4. ROC curve with threshold annotation

features), the attack by Lee et al. (CA) and the two proposed attacks GCA and ICA to both, Jsteg and Jphide. Our canonical name for the CA is $\mathrm{GCA}_{\mathrm{DC}}\ \mathrm{CR}_{\chi^2}$ with $n = 1$. There are some more specific attacks that can only detect Jsteg: the attack by Zhang and Ping (ZP) and the attack by Yu et al. (Yu). We also modified the latter two for a marginal improvement: In "$\mathrm{ZP}_{\mathrm{AC}}$" we discarded DC coefficients, since they have a different marginal distribution and "pollute" the statistics. In "Yu w/o zero bin" we excluded the zero bin from the Cauchy model approximation, since this bin shrinks when nonzero bins are growing. For the GCA and ICA we present the cases for $n = 1 \ldots 5$ for the best available measures. In general the best measure was CR_{χ^2}, except for the GCA when applied to Jphide, where we used CR_{S_2}. The overall best result was achieved for Jsteg with the GCA ($n = 2$). Next in rank follow the ICA ($n = 1$), but not significantly worse ($p = 0.07$ for bpc$= 0.04$), CA, and the 274 merged Markov and DCT features. These four attacks fulfil Ker's criterion ($\mathrm{FPR}_{0.5} \leq 0.05$) [23] for an embedding rate of 0.04 bpc. In general, the results for Jphide detection are worse for the same embedded message length. Compared to Jsteg, Jphide seems to be the better choice for steganography.

Note that there is no particular threshold that belongs to the ROC curve. Every point on the ROC curve has its own threshold that determines a false and true positive rate. To construct a classifier, we can give such a rate, and determine the threshold, which is compared to the detector output. Figure 4 shows the ROC curve for the proposed attack with the highest reliability at 0.4 bpc.

5 Conclusion

In this paper, we improved the Category Attack (CA) by Lee et al. to LSB steganography in JPEG images. By using the AC JPEG histogram, the attack was shown to be improved. Also, using the general concept of categories (GCA) and measure extensions (ICA), the attack performed with better detection reliability. It is also based only on the first order statistics, but seems to be more robust against the effect of double compression. We will study this in more detail in our ongoing work.

Acknowledgements

The work on this paper was partially supported by grant No. M10640010005-06N4001-00500 from the national R&D Program of MOST and KOSEF (first author) and by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under the research grant number FA8655-06-1-3046 (second author). The U. S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. Travel to the Information Hiding Workshop was supported in part by the European Commission through the IST Programme under contract IST-2002-507932 ECRYPT.

References

1. Upham, D.: Jsteg (1993) Online available at `http://munitions.vipul.net/software/steganography/jpeg-jsteg-v4.diff.gz`
2. Westfeld, A., Pfitzmann, A.: Attacks on steganographic systems. In Pfitzmann, A., ed.: Information Hiding (3rd International Workshop). Volume 1768 of LNCS., Berlin Heidelberg, Springer-Verlag (2000) 61–76
3. Provos, N.: Outguess (2001) Online available at `http://www.outguess.org`
4. Westfeld, A.: F5—a steganographic algorithm: High capacity despite better steganalysis. In Moskowitz, I.S., ed.: Information Hiding (4th International Workshop). Volume 2137 of LNCS., Berlin Heidelberg, Springer-Verlag (2001) 289–302
5. Sallee, P.: Model-based steganography. In Kalker, T., Ro, Y.M., Cox, I.J., eds.: International Workshop on Digital Watermarking. Volume 2939 of LNCS., Berlin Heidelberg, Springer-Verlag (2004) 154–167
6. Shi, Y.Q., Chen, C., Chen, W.: A Markov process based approach to effective attacking JPEG steganography. In Camenisch, J., Collberg, C., Johnson, N.F., Sallee, P., eds.: Information Hiding (8th International Workshop). Volume 4437 of LNCS., Berlin Heidelberg, Springer-Verlag (2007)
7. Pevný, T., Fridrich, J.: Merging Markov and DCT features for multi-class JPEG steganalysis. In Delp III, E.J., Wong, P.W., eds.: Security, Steganography and Watermarking of Multimedia Contents IX (Proc. of SPIE), San Jose, CA (2007)
8. Fridrich, J., Goljan, M., Soukal, D.: Higher-order statistical steganalysis of palette images. In Delp III, E.J., Wong, P.W., eds.: Security, Steganography and Watermarking of Multimedia Contents V (Proc. of SPIE), San Jose, CA (2003) 178–190

9. Fridrich, J., Goljan, M., Du, R.: Detecting LSB steganography in color and grayscale images. IEEE Multimedia **8**(4) (2001) 22–28
10. Dumitrescu, S., Wu, X., Wang, Z.: Detection of LSB steganography via sample pair analysis. In Petitcolas, F.A.P., ed.: Information Hiding (5th International Workshop). Volume 2578 of LNCS., Berlin Heidelberg, Springer-Verlag (2003) 355–372
11. Westfeld, A.: Space filling curves in steganalysis. In Delp III, E.J., Wong, P.W., eds.: Security, Steganography and Watermarking of Multimedia Contents VII (Proc. of SPIE), San Jose, CA (2005) 28–37
12. Lu, P., Luo, X., Tang, Q., Shen, L.: An improved sample pairs method for detection of LSB embedding. In Fridrich, J., ed.: Information Hiding (6th International Workshop). Volume 3200 of LNCS., Berlin Heidelberg, Springer-Verlag (2004) 116–127
13. Yu, X., Wang, Y., Tan, T.: On estimation of secret message length in Jsteg-like steganography. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04). (2004) 673–676
14. Zhang, T., Ping, X.: A fast and effective steganalytic technique against Jsteg-like algorithms. In: Proceedings of the 2003 ACM Symposium on Applied Computing (SAC2003), March 9–12, 2003, Melbourne, Florida, USA, New York, ACM Press (2003) 307–311
15. Lee, K., Westfeld, A., Lee, S.: Category Attack for LSB steganalysis of JPEG images. In Shi, Y.Q., Jeon, B., eds.: Digital Watermarking (5th International Workshop) IWDW 2006 Jeju Island, Korea, November 8–10, 2006, Revised Papers. Volume 4283 of LNCS., Berlin Heidelberg, Springer-Verlag (2006) 35–48
16. Westfeld, A.: Steganalysis in the presence of weak cryptography and encoding. In Shi, Y.Q., Jeon, B., eds.: Digital Watermarking (5th International Workshop) IWDW 2006 Jeju Island, Korea, November 8–10, 2006, Revised Papers. Volume 4283 of LNCS., Berlin Heidelberg, Springer-Verlag (2006) 19–34
17. University of Washington: CBIR image database (2004) Online available at `http://www.cs.washington.edu/research/imagedatabase/groundtruth`
18. NRCS: Photo gallery of the USDA Natural Resources Conservation Service (2006) Online available at `http://photogallery.nrcs.usda.gov/`
19. Smoot, S.R., Rowe, L.A.: Study of DCT coefficient distributions. In: Proceedings of the SPIE Symposium on Electronic Imaging. Volume 2657., San Jose, CA (1996) 403–411
20. Fridrich, J., Goljan, M., Hogea, D.: Steganalysis of JPEG images: Breaking the F5 algorithm. In Petitcolas, F.A.P., ed.: Information Hiding (5th International Workshop). Volume 2578 of LNCS., Berlin Heidelberg, Springer-Verlag (2003) 310–323
21. Provos, N., Honeyman, P.: Detecting steganographic content on the Internet (2001) Online available at `http://www.citi.umich.edu/techreports/reports/citi-tr-01-11.pdf`
22. Fridrich, J.: Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes. In Fridrich, J., ed.: Information Hiding (6th International Workshop). Volume 3200 of LNCS., Berlin Heidelberg, Springer-Verlag (2004) 67–81
23. Ker, A.D.: Improved detection of LSB steganography in grayscale images. In Fridrich, J., ed.: Information Hiding (6th International Workshop). Volume 3200 of LNCS., Berlin Heidelberg, Springer-Verlag (2004) 97–115