Targeted Attacks on Quantization-based Watermarking Schemes

Peter Meerwald, Christian Koidl, Andreas Uhl

September 16, 2009

Overview

- What are *targeted attacks*?
- Attack targets and examplatory attack
- Results and conclusions

Targeted Attacks

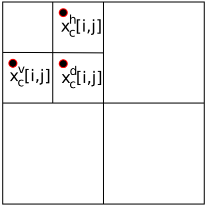
- Targeted attacks assume full knowledge about the watermarking scheme except the key (Kerckhoffs' principle [Kerckhoffs, 1883]).
- Consider watermark-only-attack (WOA): want to remove watermark with access to only a *single* watermarked image.
- We do not discuss robustness attacks (signal processing, compression) here, but *watermark security.*
- Watermark security "refers to the inability of an unauthorized user to have access to the raw watermarking channel" [Kalker, 2001].

Attack Targets

- Quantization of Middle Wavelet Detail Coefficients (QMWDC) [Kundur and Hatzinakos, 1998]
- Wavelet-Tree Quantization (WTQ) [Wang and Lin, 2004]
- Structure-Based Wavelet Tree Quantization (SBWTQ) [Wu and Huang, 2007]
- Watermarking Technique based on JPEG2000 Codec (WTJC) [Chen et al., 2004]
- Double Wavelet Tree Energy Modulation (DWTEM) [Tsai et al., 2008]
- Significant Difference of Wavelet Coefficient Quantization (SDWCQ) [Lin et al., 2008]

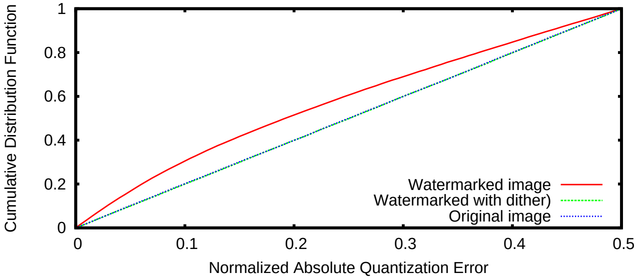
Analysis of QMWDC (1)

- Quantization of Middle Wavelet Detail Coefficients (QMWDC) embeds a binary watermark in wavelet-domain detail subband coefficients.
- A secret key selects embedding positions with coefficient triples $(x_c^h[i,j], x_c^v[i,j], x_c^d[i,j])$.
- The coefficients of each triple are ordered (x^s, x^m, x^l) where $x^s \leq x^m \leq x^l$ and the middle coefficient x^m is quantized using bin width $\Delta_c = {}^{(x_c^l - x_c^s)}/_{(2Q-1)}$ to embed one watermark bit.



Analysis of QMWDC (2)

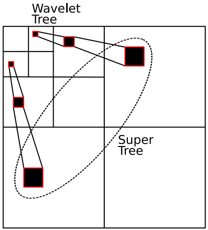
- The absolute quantization error $e_c = |\text{rnd}(x_c^m/\Delta_c) - x_c^m/\Delta_c|$ normalized by the corresponding quantization bin width for each possible embedding position $[i,j]$ shows a clear bias towardsw smaller errors in the CDF for the watermarked image.



- The bias allows to estimate embedding positions, $\Delta_c[i,j]$ reveals the optimal attack power.
- Countermeasure: dither vector prevents estimation of embedding positions.

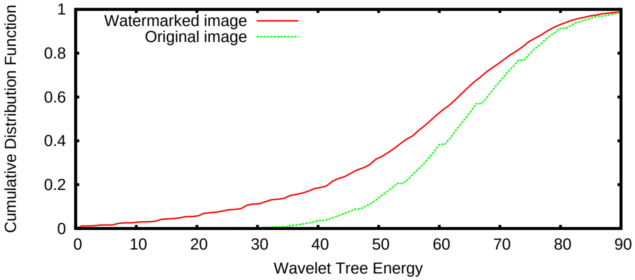
Analysis of WTQ (1)

- Wavelet-Tree Quantization (WTQ) quantizes coefficients of a wavelet tree.
- Several trees are randomly selected and combined into super-trees to embed one bit.



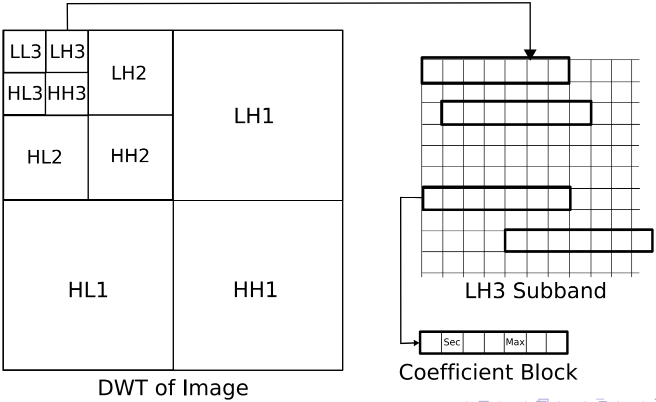
Analysis of WTQ (2)

- WTQ permutes the order of wavelet tree to disguise the relation of wavelet-tree to super-tree.
- However, coefficients belonging to one wavelet tree are known.
- The energy of quantized wavelet trees differs significantly from non-quantized trees allowing to guess the embedding locations.



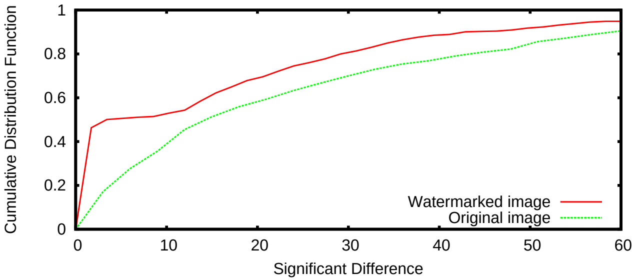
Analysis of SDWCQ (1)

- Significant Difference Wavelet Coefficient Quantization (SDWCQ) groups several adjacent coefficients into a blocks which are shuffled.
- Within each block, the significant difference d between the largest and second largest coefficient, max and sec, is made large to encode 1 and small to encode -1.



Analysis of SDWCQ (2)

- The shuffling only encrypts the watermark message but does not protect the watermark channel.
- The CDF of significant differences for all possible blocks differs noticeable.



- Countermeasure: Shuffle coefficients before block formation so that significant difference can not be computed; scheme is still not secure.

Lessons Learnt and Improvements

- Many wavelet-domain quantization-based watermarking schemes leak information allowing to mount an efficient attack.
- The attack methods are related to targeted steganalysis (analysis of statistics).
- The structures employed (wavelet trees, coefficient blocks, subbands) facilitate the attack.
- Security measures (permutation, dithering) are insufficient or missing altogether.
- Watermarking for copyright protection application requires robustness <u>and</u> security.

Experimental Results

- Present attack results on ten 512×512 grayscale images, separate WOA
- Normalized Correlation (NC) measure for watermark strength
- Image quality in PSNR (dB)
  - for watermarked image against attacked image (w, a)
  - for original image against attacked image (o, a)
  - for original image against watermarked image (o, w)

QMWDC Attack Results

Image	∅ NC	∅ PSNR (dB)		
		(w,a)	(o,a)	(o,w)
Lena	0.021	54.29	45.79	46.13
Goldhill	0.014	52.36	44.99	45.42
Peppers	0.056	54.64	45.31	45.61
Man	0.039	51.57	43.01	43.29
Airport	0.064	51.02	42.22	42.48
Tank	−0.009	53.01	47.46	48.18
Truck	−0.032	52.97	47.00	47.62
Elaine	0.073	53.55	47.17	47.79
Boat	−0.036	52.28	43.39	43.69
Barbara	−0.063	50.80	42.54	42.83
Average	0.013	52.65	44.89	45.30

Image	∅ NC	∅ PSNR (dB)		
		(w,a)	(o,a)	(o,w)
Lena	0.028	50.05	45.06	46.11
Goldhill	−0.054	48.54	44.15	45.32
Peppers	−0.018	51.02	44.73	45.49
Man	−0.005	47.21	42.27	43.24
Airport	0.009	47.84	41.73	42.48
Tank	−0.037	50.34	46.71	48.17
Truck	−0.023	49.62	46.06	47.52
Elaine	−0.043	51.04	46.58	47.76
Boat	−0.012	48.66	42.87	43.70
Barbara	0.018	48.27	41.99	42.71
Average	−0.013	49.26	44.22	45.25

without and with dither quantization, $Q = 4$

WTJC Attack Results

Image	∅ NC	∅ PSNR (dB) (w,a)	(o,a)	(o,w)
Lena	−0.007	47.18	39.74	40.30
Goldhill	−0.024	47.95	41.02	41.68
Peppers	0.023	48.10	40.38	40.88
Man	0.118	50.57	41.56	41.92
Airport	0.048	49.55	42.43	43.02
Tank	−0.152	42.81	39.11	41.27
Truck	0.071	48.83	39.70	40.02
Elaine	−0.029	46.25	39.19	39.82
Boat	−0.073	45.73	39.63	40.55
Barbara	−0.021	47.46	40.36	40.98
Average	−0.005	47.44	40.31	41.04

$\alpha = 0.6$ with distortion reduction

SBWTQ Attack Results

Image	∅ NC	∅ PSNR (dB)		
		(w,a)	(o,a)	(o,w)
Lena	0.000	54.76	44.57	44.73
Goldhill	0.000	51.15	42.12	41.31
Peppers	0.000	53.49	41.40	41.38
Man	0.000	51.95	42.02	41.68
Airport	0.000	51.14	41.37	40.92
Tank	0.000	51.24	44.63	44.08
Truck	0.000	50.66	42.27	41.53
Elaine	0.000	53.08	44.90	44.87
Boat	0.000	54.17	42.43	42.46
Barbara	0.000	53.03	42.77	42.56
Average	0.000	52.47	42.85	42.55

$$\Delta = 10$$

WTQ Attack Results

Image	⌀ NC	⌀ PSNR (dB)		
		(w,a)	(o,a)	(o,w)
Lena	−0.049	49.55	40.90	41.49
Goldhill	0.063	51.13	44.92	45.82
Peppers	−0.121	49.83	43.51	44.54
Man	0.122	51.52	45.49	46.30
Airport	0.116	51.89	45.93	46.81
Tank	−0.036	51.54	46.22	47.24
Truck	0.002	51.20	45.80	46.85
Elaine	−0.177	50.31	45.29	46.68
Boat	0.023	50.63	43.39	44.12
Barbara	0.073	50.45	42.51	43.11
Average	0.001	50.81	44.40	45.30

$E = 100, q_{max} = 336$ and $\epsilon = 0.1$

DWTEM Attack Results

Image	∅ NC	∅ PSNR (dB)		
		(w,a)	(o,a)	(o,w)
Lena	0.228	44.93	39.77	41.08
Goldhill	0.222	42.44	39.60	41.90
Peppers	0.217	43.94	40.07	41.92
Man	0.229	39.07	36.75	39.38
Airport	0.229	38.24	36.63	39.92
Tank	0.222	44.99	43.39	47.16
Truck	0.225	43.23	41.40	44.80
Elaine	0.225	45.18	41.89	44.31
Boat	0.224	38.06	36.04	39.54
Barbara	0.229	36.90	35.23	39.35
Average	0.225	41.70	39.08	41.93

$$\Delta = 0.15$$

SDWCQ Attack Results

Image	∅ NC	∅ PSNR (dB)		
		(w,a)	(o,a)	(o,w)
Lena	0.020	54.42	46.42	46.63
Goldhill	−0.109	53.36	45.79	45.91
Peppers	−0.023	54.08	45.02	45.05
Man	0.025	51.94	42.70	42.85
Airport	−0.108	53.00	45.00	45.10
Tank	−0.112	54.22	48.81	48.97
Truck	−0.121	52.43	44.79	44.96
Elaine	−0.066	54.39	47.01	47.37
Boat	−0.040	53.79	45.69	45.82
Barbara	−0.014	53.96	46.04	46.19
Average	−0.055	53.56	45.73	45.88

γ unrestrained, block size 7, $T = 12$ and $\alpha = 0.9$

Conclusion

- Several quantization based watermarking schemes for copyright protection in the wavelet domain have been shown insecure
- Wavelet-tree structure exposes too much structure for attack
- Many more proposals likely vulnerable
- Security issue is often ignored, no security measures implemented
- Source code available at `http://www.wavelab.at/sources`

Chen, T.-S., Chen, J., and Chen, J.-G. (2004).
A simple and efficient watermark technique based on JPEG2000 codec.
ACM Multimedia Systems Journal, 10(1):16–26.

Kalker, T. (2001).
Considerations on watermarking security.
In *Proceedings of the IEEE Workshop on Multimedia Signal Processing, MMSP '01*, pages 201–206, Cannes, France.

Kerckhoffs, A. (1883).
La cryptographie militaire.
Journal des sciences militaires, 9:5–83.

Kundur, D. and Hatzinakos, D. (1998).
Digital watermarking using multiresolution wavelet decomposition.
In *Proceedings of the 1998 International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, volume 5, pages 2969–2972, Seattle, WA, USA.

Lin, W.-H., Horng, S.-J., Kao, T.-W., Fan, P., Lee, C.-L., and Pan, Y. (2008).
An efficient watermarking method based on significant difference of wavelet coefficient quantization.
IEEE Transactions on Multimedia, 10(5):746–757.

Tsai, M.-J., Lin, C.-T., and Liu, J. (2008).
A wavelet-based watermarking scheme using double wavelet tree energy modulation.
In *Proceedings of the 2008 IEEE International Conference on Image Processing, ICIP '08*, pages 417–420, San Diego, CA, USA. IEEE.

Wang, S.-H. and Lin, Y.-P. (2004).
Wavelet tree quantization for copyright protection watermarking.
IEEE Transactions on Image Processing, 13(2):154–165.

Wu, G.-D. and Huang, P.-H. (2007).
Image watermarking using structure based wavelet tree quantization.
In *Proceedings of the 6th IEEE/ACIS International Conference on Computer and Information Science, 2007. ICIS 2007*, pages 315–319. IEEE.