

On Dangers of Overtraining Steganography to Incomplete Cover Model

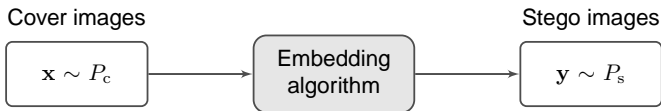
Jan Kodovský, Jessica Fridrich, Vojtěch Holub

September 30, 2011 / MMSEC



Steganography

- Steganography by cover modification



- $\mathbf{x}, \mathbf{y} \in \mathcal{C} \dots$ space of images, e.g. $\mathcal{C} = \{0, \dots, 255\}^{512 \times 512}$
- Goal: keep P_s close to P_c (minimize KL divergence)
- Steganographer's options:
 - Map everything into feature space \mathcal{F} and preserve cover pdf there
 - Minimize distortion function $D(\mathbf{x}, \mathbf{y})$

OutGuess

- JPEG domain steganographic algorithm [Provos 2001]
- Feature space \mathcal{F} ... histogram of DCT coefficients
- Statistical restoration (LSB embedding + correction)
- Fully preserves cover pdf in $\mathcal{F} \Rightarrow$ undetectable within \mathcal{F}
- Problem: first order statistics is a poor model of cover images
- Successful attacks using higher order statistics [2002-2011]

Overtrained to incomplete model

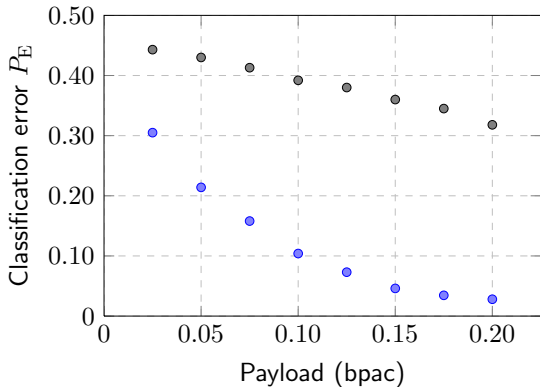
Feature Correction Method (FCM)

- General framework for steganography that approximately preserves given feature vector [Kodovský 2008, Chonev 2011]
- Minimize distortion function defined in \mathcal{F} :

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \frac{(x_i - y_i)^2}{var_i} \quad \begin{array}{l} n \dots \text{number of features} \\ var_i \dots \text{variance of the } i\text{th feature} \end{array}$$

- Two-pass procedure, greedy approach, already in the first phase make ± 1 changes to minimize distortion
- Implemented for the case of 274 PEV features [Pevný 2007]
 - Captures both inter- and intra-block dependencies
 - Local histograms, co-occurrences, Markov models

Feature Correction Method (FCM)



- Performance in \mathcal{F}
- Different cropping

$\mathcal{F} \equiv 274$ PEV

SVM classifier

$$P_E = \min_{P_{FA}} \frac{P_{FA} + P_{MD}}{2}$$

Overtrained to incomplete model

Optimized ± 1 embedding in JPEG domain

- Minimal-distortion steganography [Filler 2011]
- Adaptive scheme with empirically designed distortion function:

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^N \rho_i(\mathbf{x}, y_i), \quad N \dots \text{number of changeable coefficients}$$

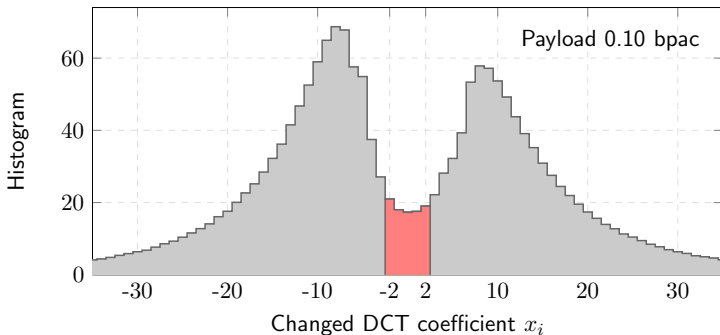
where $\rho_i(\mathbf{x}, y_i) \in \mathbb{R}$ is the cost of changing $x_i \rightarrow y_i$

- Costs are functions of inter- and intra-block neighbors optimized w.r.t. given model (feature space)
- Will be abbreviated MOD (**M**odel **O**ptimized **D**istortion)

Details of MOD algorithm

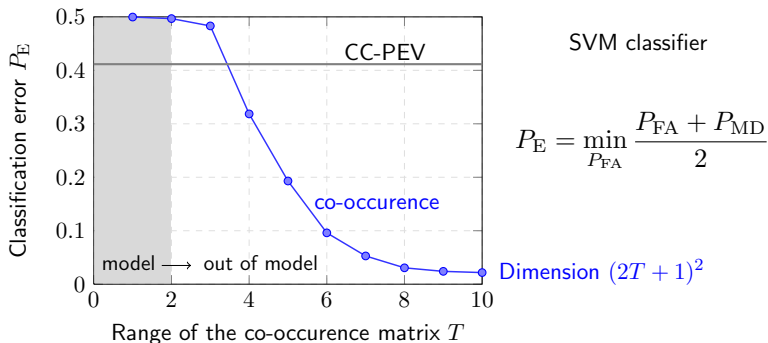
- Optimize parameters of the cost function
 - To **minimize** the margin of linear SVM in \mathcal{F}
 - Nelder-Mead simplex-reflection algorithm
 - Fixed misclassification cost C
 - Less than 100 images needed
- Feature space $\mathcal{F} \equiv 548$ -dim CC-PEV [Kodovský 2009]
 - CC-PEV = PEV enhanced by Cartesian calibration
- Preliminary experiments showed no indication of overtraining
 - Different calibration cropping
 - CDF = CC-PEV + SPAM features [Pevný 2009]

Histogram of changed DCT coefficients



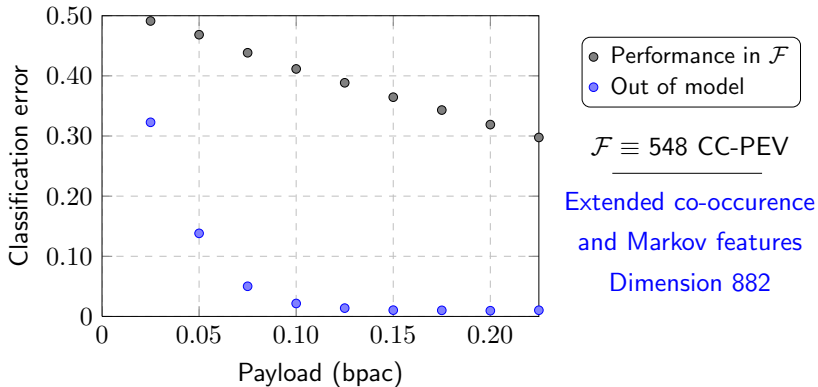
- CC-PEV: inter-block co-occurrences are constrained to $[-2,2]$
- 95% of changes are made *out of the model*

Extending the model



- Extending inter-block co-occurrences compromises the security
- We can extend the range of other parts of the CC-PEV model

Attacking MOD algorithm



Optimization moved changes out of the incomplete model

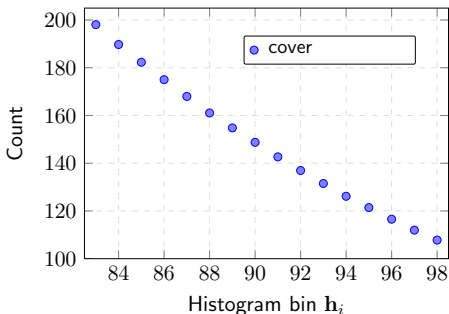
HUGO

- Minimal distortion steganography in spatial domain [Pevný 2010]
- BOSS (Break Our Steganographic System)
- \mathcal{F} ... 3D co-occurrence matrices of pixel differences
 - For differences $\mathbf{d} = (d_1, d_2, d_3)$... co-occurrence bin value $\mathbf{C}_{\mathbf{d}}(\mathbf{x})$
 - $T = 90 \Rightarrow$ dimension $2(2T + 1)^3 = 11,859,482$
- Distortion function - weighted L_1 -norm in \mathcal{F}

$$D(\mathbf{x}, \mathbf{y}) = \sum_{d_1, d_2, d_3 = -T}^T \frac{1}{1 + \|\mathbf{d}\|_2} \cdot |\mathbf{C}_{\mathbf{d}}(\mathbf{x}) - \mathbf{C}_{\mathbf{d}}(\mathbf{y})|$$

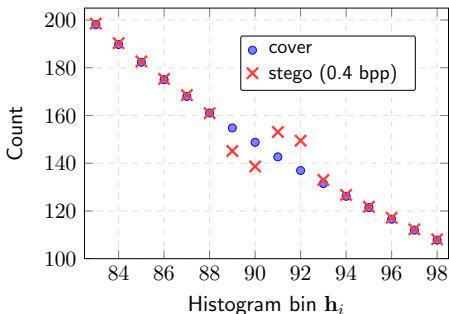
Model weakness

- Abrupt end at $T = 90 \Rightarrow$ model treats pixels above T differently
- 3D co-occurrence is sparse around $T = 90 \Rightarrow$ form *marginals*
- $\mathbf{h}_i(\mathbf{x})$... number of adjacent pixel pairs whose difference is i



Model weakness

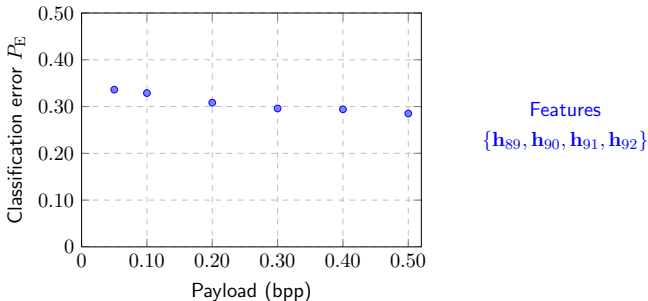
- Abrupt end at $T = 90 \Rightarrow$ model treats pixels above T differently
- 3D co-occurrence is sparse around $T = 90 \Rightarrow$ form *marginals*
- $h_i(\mathbf{x}) \dots$ number of adjacent pixel pairs whose difference is i



Attacking HUGO with 4 features

- Almost payload-independent detection
- Works better on noisy and textured images

Abrupt end of model creates security weakness



Conclusions

- Overtraining to simplified cover model seems to be a common security flaw of modern schemes
- It is difficult to find a good (complete) model for cover images
- Steganography: high dimension is not sufficient
Steganalysis: high dimension is not necessary
- Straightforward security improvements:
 - ✓ HUGO ... increase $T = 255$
 - ? MOD embedding ... increase range in CC-PEV
- Suggestion: use rich and diverse models