On Completeness of Feature Spaces in Blind Steganalysis

Jan Kodovský Jessica Fridrich

September 23 / MM&SEC 2008

BINGHAMTON
UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Feature Spaces in Blind Steganalysis

- In blind steganalysis, the feature set plays the role of a low-dimensional *image model*.
- Good low-dimensional models are used for
  - Steganalysis
  - Benchmarking
  - Design of steganographic schemes (blind steganalyzer used as an oracle)
- For these applications, it is important that the features completely describe natural images
  - e.g., if a stego method preserves the whole feature vector, it should be undetectable using other features.

Motivation

- Notation:

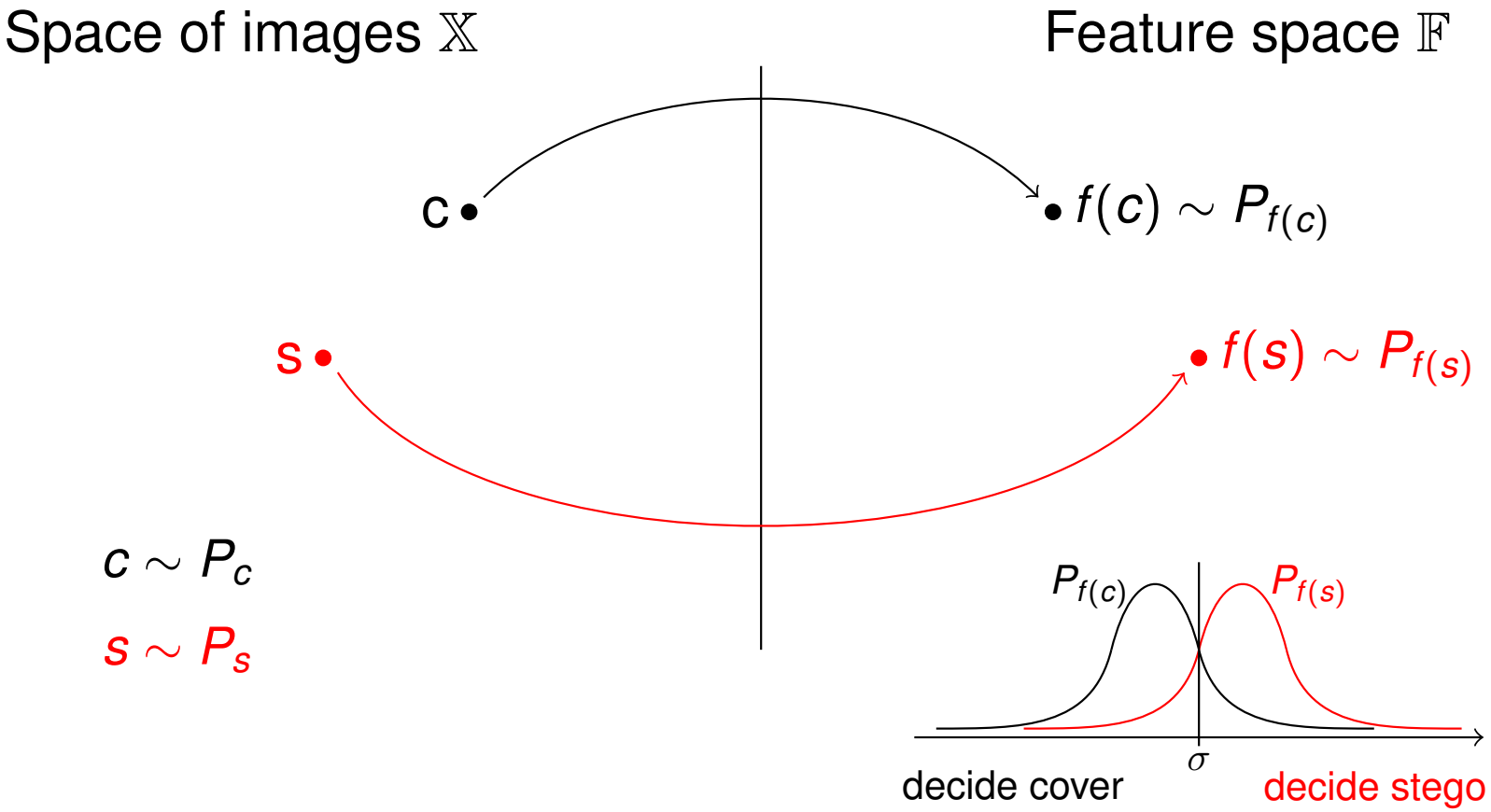
  $\mathbb{X}$ ... original space of images,
  e.g., $\mathbb{X} = \{0, \ldots, 255\}^{N \times N}$

  $\mathbb{F}$ ... low-dimensional feature space

  f ... feature map, $f : \mathbb{X} \to \mathbb{F}$

- Our goals:
  - Decide whether or not a given feature space $\mathbb{F}$ completely describes cover images
  - Ability to refute *completeness* experimentally

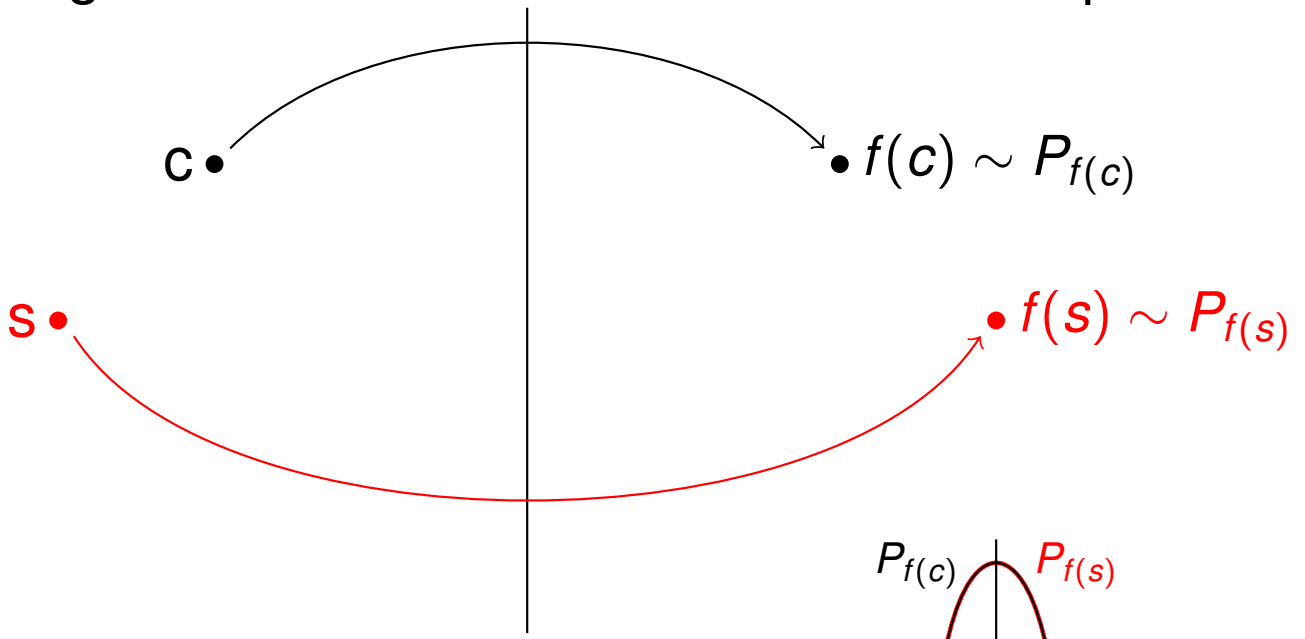
Motivation


Space of images $\mathbb{X}$
Feature space $\mathbb{F}$
c
$f(c) \sim P_{f(c)}$
s
$f(s) \sim P_{f(s)}$
$c \sim P_c$
$s \sim P_s$
$P_{f(c)}$
$P_{f(s)}$
σ
decide cover
decide stego

Motivation

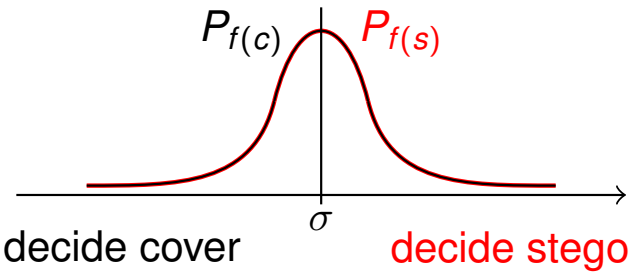
Space of images $\mathbb{X}$

Feature space $\mathbb{F}$



$c \sim P_c$

$s \sim P_s$



- What if $P_{f(c)} = P_{f(s)}$?

$\Rightarrow$ undetectability within the given feature space

Motivation

Two possibilities:

- $P_c = P_s$ in the original space $\mathbb{X}$
  - perfect steganography (unlikely)
- $P_c \neq P_s$ in the original space $\mathbb{X}$
  - feature space $\mathbb{F}$ is not a complete descriptor of cover images
  - there exists a different feature space $\mathbb{F}'$ in which $P_{f'(c)} \neq P_{f'(s)}$ (at least in theory)

Proposed Approach

- Given the feature space $\mathbb{F}$, we construct a steganographic method that approximately preserves the feature vector
  $\Rightarrow P_{f(c)} \approx P_{f(s)}$
  Feature Correction Method (FCM)
- If we find a different space $\mathbb{F}'$ in which the proposed method is detectable $\Rightarrow$ feature space $\mathbb{F}$ is not a complete descriptor of cover images

Feature Correction Method (FCM)

- FCM approximately preserves the entire feature vector.
- Embedding procedure
  - Split the set of all DCT coefficients $\mathcal{D}$ into $\mathcal{D}_e \cup \mathcal{D}_c$.
  - Embed payload in non-zero coefficients from $\mathcal{D}_e$ by modifying them by ± 1 while choosing the direction that perturbs the feature vector the least (requires WPCs).
  - Use DCTs from $\mathcal{D}_c$ to reduce the final distortion even more, using changes by ± 1 and ± 2.

Feature Correction Method (FCM)

- How to measure distance in feature space?

  $\rightarrow \quad d(x,y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{var_i}$, var_i . . . variance of i-th feature on covers

- How to split $\mathcal{D}$ into $\mathcal{D}_e$ and $\mathcal{D}_c$?

  $\rightarrow$ Experimentally

- Is it computationally realisable?

  $\rightarrow$ Differential feature computation

Feature Set Used in Experiments

- 274 Merged extended DCT and Markov features
  - Global histograms (11)
  - 5 local AC histograms (5×11)
  - 11 dual histograms (11×9)
  - Variation (1)
  - Blockiness (2)
  - Co-occurence matrix (25)
  - Markov features (81)

  } 193 extended DCT features

+ calibration

(Pevný et al., SPIE 2007)

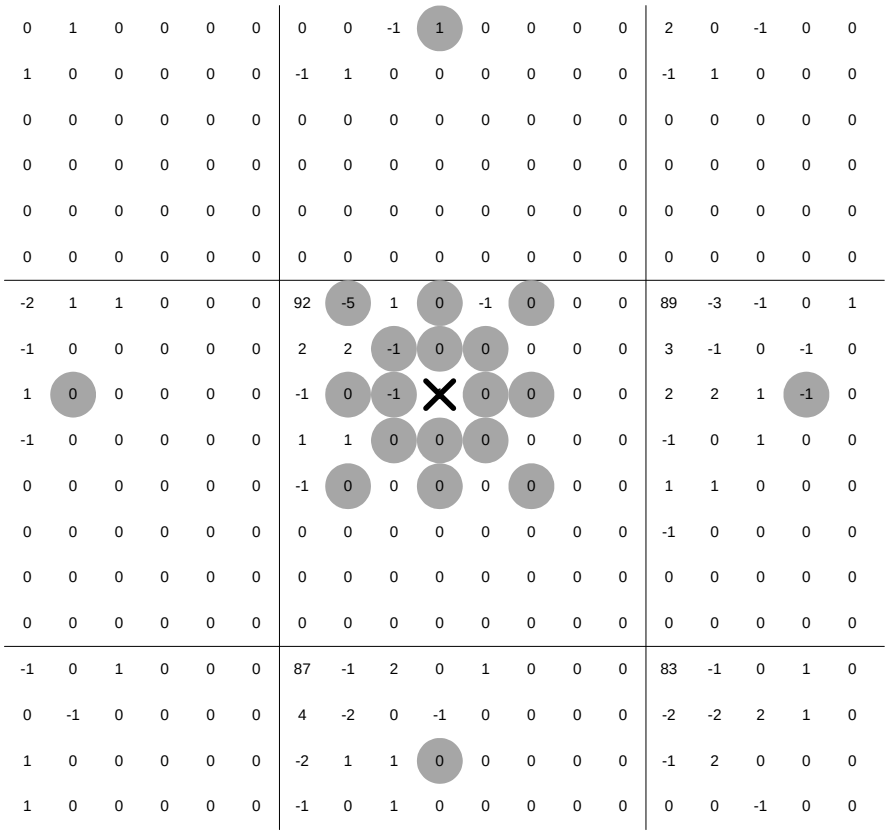
Differential Feature Computation

- Calculation of feature vector is $O(N)$, where N is number of DCT coefficients
- We need to update feature vector after every DCT flip
- Recalculating every time $\rightarrow O(N^2)$ ..... infeasible
- Solution: differential feature computation $\rightarrow O(N)$
- Example (global DCT histogram):
  - modify DCT coefficient value from d to $d+1$:

$$\begin{aligned} h[d] &\leftarrow h[d]-1 \\ h[d+1] &\leftarrow h[d+1]+1 \end{aligned}$$

Differential Feature Computation

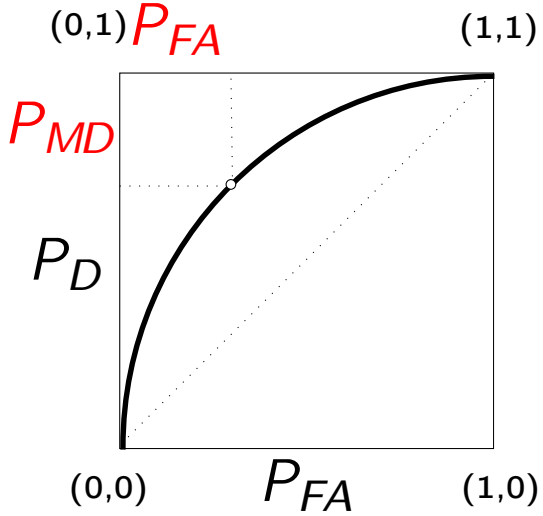
- Higher order statistics, Markov features:



Evaluating Security

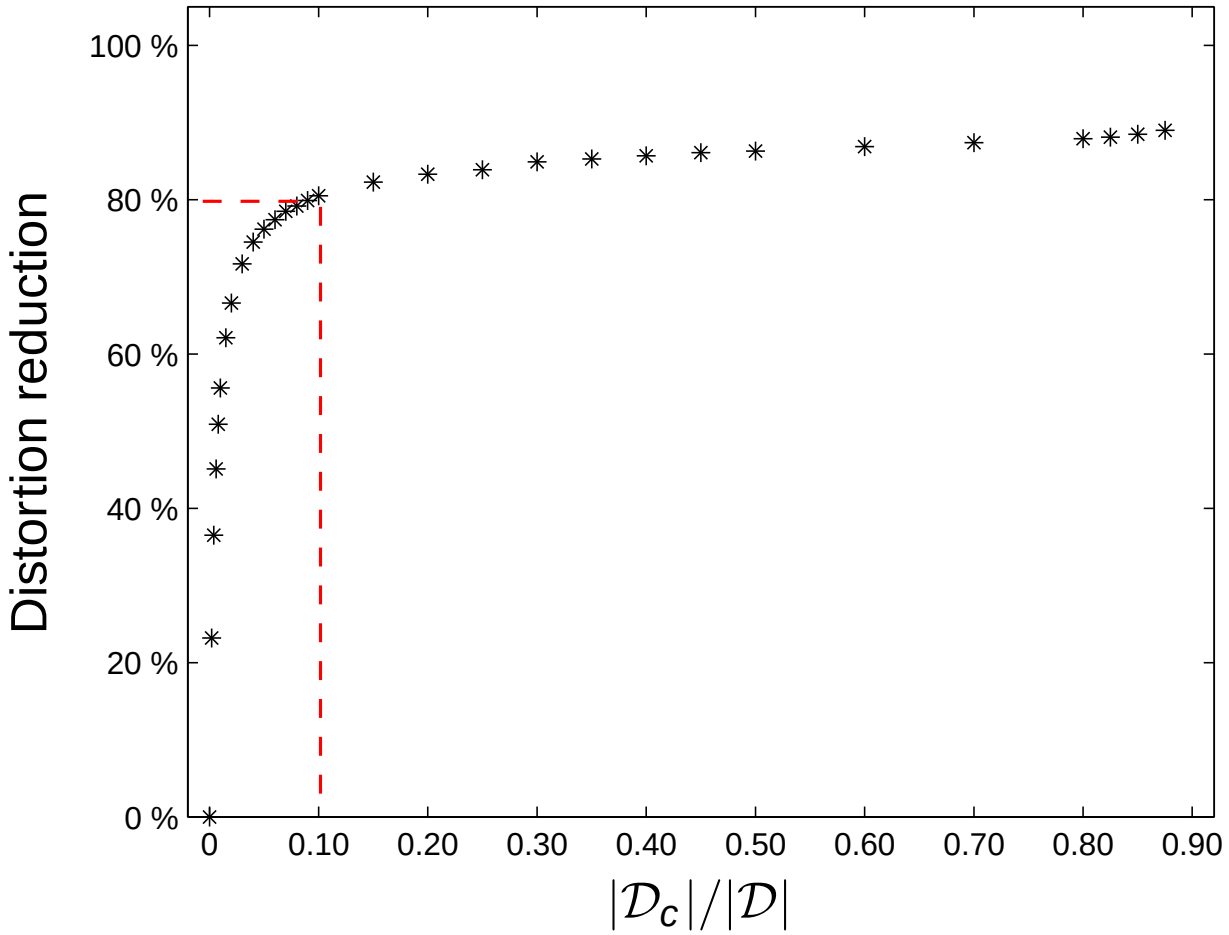
- Blind steganalyzer (Pevný et al., SPIE 2007)
  - SVM machine with Gaussian kernel
  - 6000 images, single compressed 75% JPEGs, smaller side 512 pixels, grayscale
  - 3500 training and 2500 testing images
- Detection error

$$P_E = \min \frac{1}{2}(P_{FA} + P_{MD})$$



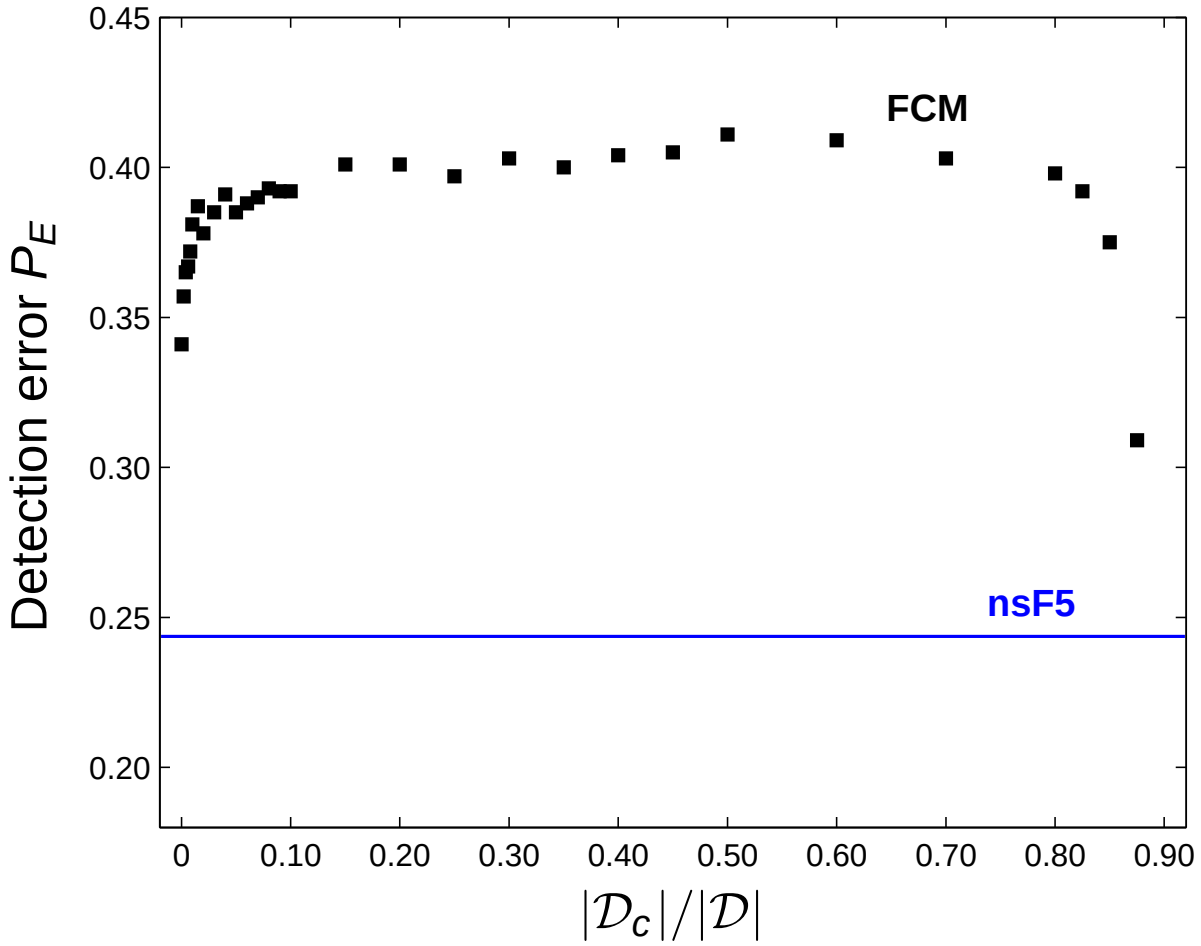
Experimental Results

- Distortion reduction (0.10 bpac, avg. over 6,000 images)



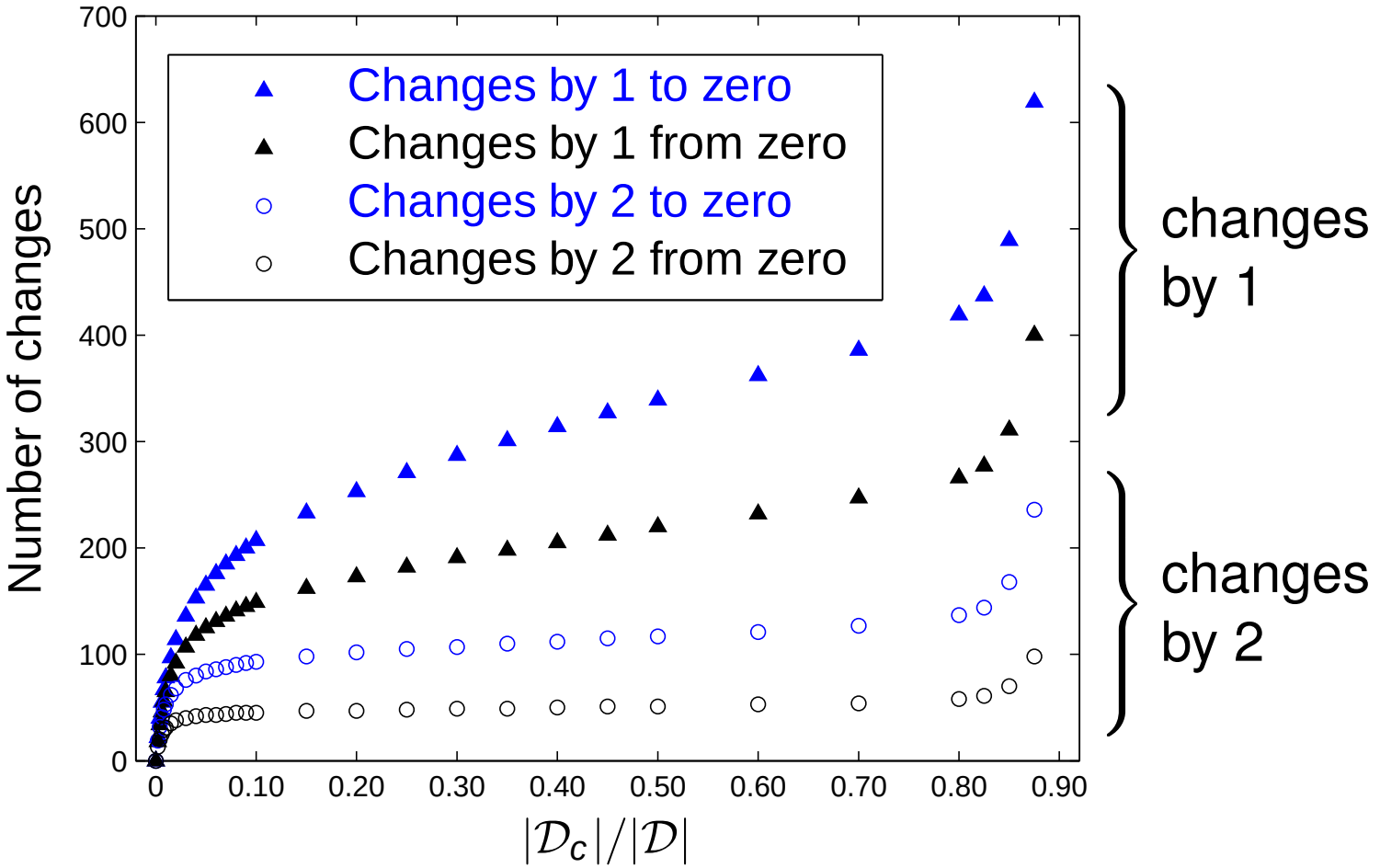
Experimental Results

- Detection error P_E (payload 0.10 bpac)



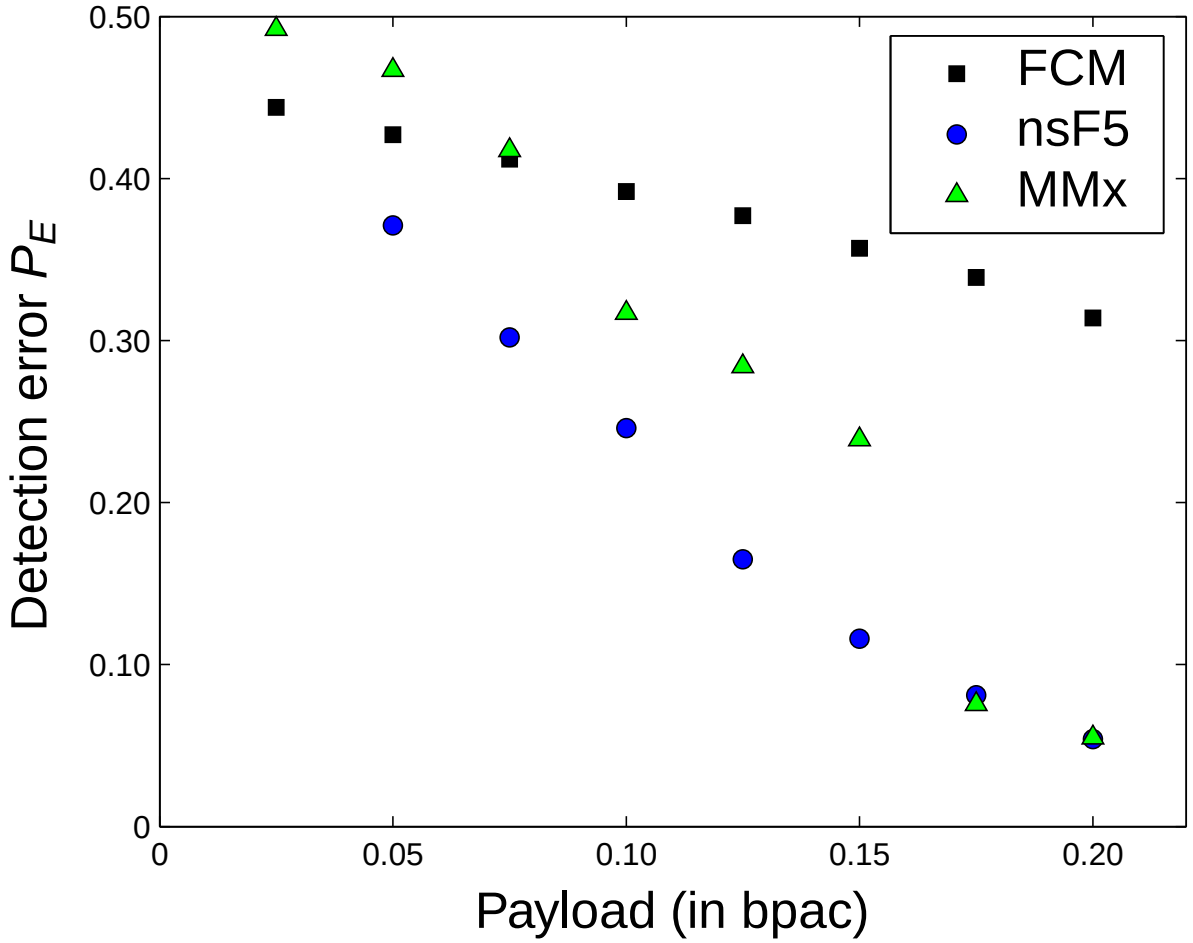
Experimental Results

- Character of corrections in $\mathcal{D}_c$ (payload 0.10 bpac)



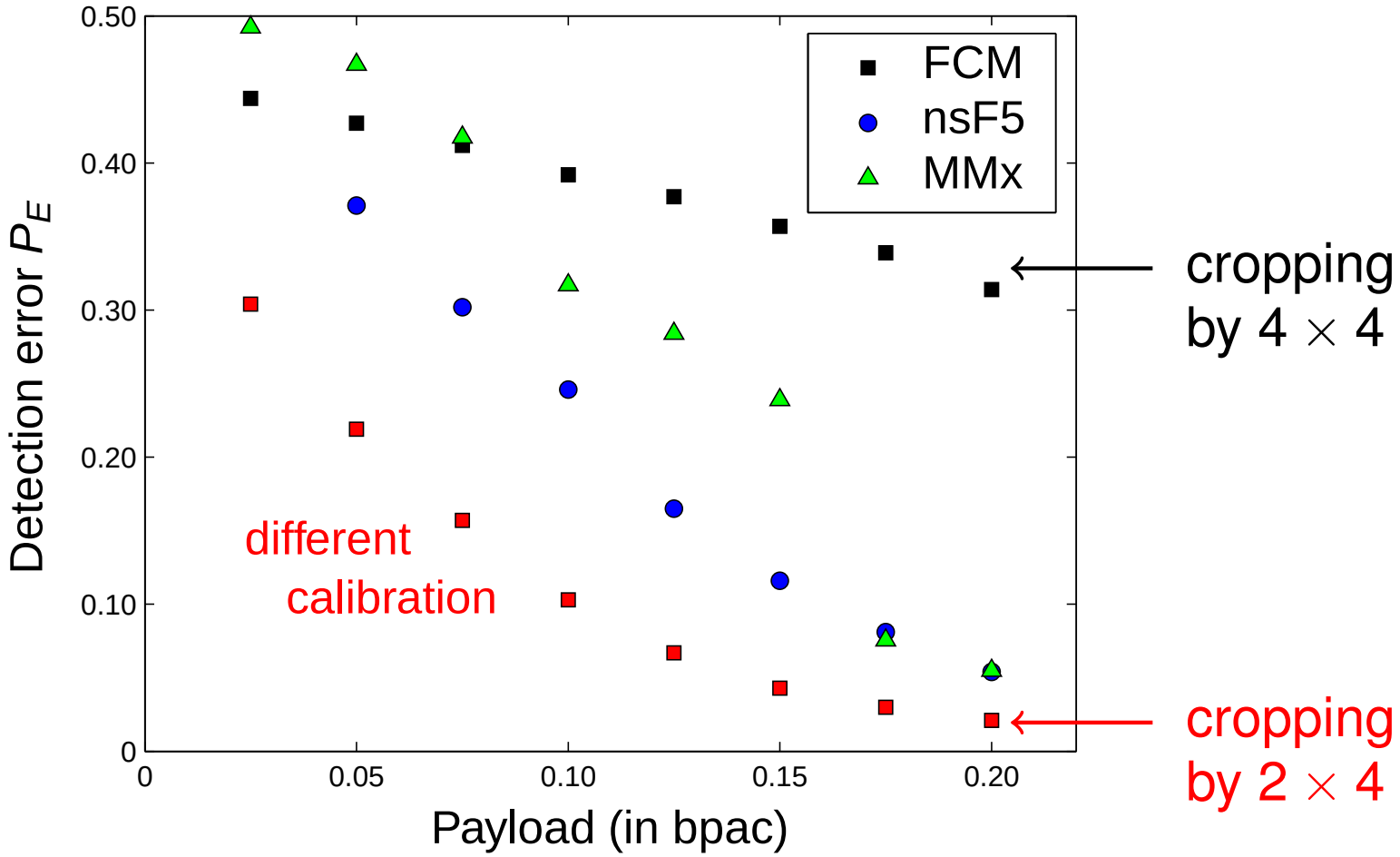
Experimental Results

- Comparison with nsF5 and MMx (for $|\mathcal{D}_c|/|\mathcal{D}| = 0.10$)



Experimental Results

- Comparison with nsF5 and MMx (for $|\mathcal{D}_c|/|\mathcal{D}| = 0.10$)



Summary

- Many applications require low dimensional image models to be complete.
- Proposed the concept of completeness that is experimentally refutable.
- Feature Correction Method (FCM) is a steganographic method that approximately preserves the whole feature vector.
  - FCM is undetectable within a given image model (feature space).
  - If FCM is detectable using a different feature space, the original feature space is incomplete and can be augmented.