Generalized Transfer Component Analysis for Mismatched Jpeg Steganalysis

Xiaofeng Li[1], Xiangwei Kong[1], Bo Wang[1], Yanqing Guo[1],Xingang You[2]

[1]School of Information and Communication Engineering
Dalian University of Technology, Dalian, 116024, China

[2]Beijing Institute of Electronic Technology and Application
Beijing, 100091, China

Contents

Contents

Motivation of Our Work

Battle of steganography and steganalysis

Steganography

- embed message signal into cover images to get stego images;
- message undetectable in covert communication

steganalysis

- features sensitive to change due to embedding
- build decision model using machine learning
- recognize stego images from plain cover images

- Steganalysis seem to win the battle recently
  [Kodovsky and Fridrich 12]: Rich model perform well to detect six modern steganographic schemes at low embedding rate.

Motivation of Our Work

Steganalysis Really Win the Battle?

Success of state of the art steganalysis methods rely on having prior knowledge of steganography to build the training set.
——cover images, embedding algorithm is known.

Matched steganalysis

train set and test set:
- ☐ **matched cover images**
- ☐ **matched embedding algorithm**

laboratory

Mismatched steganalysis

train set and test set:
- ☐ **mismatched cover images**
- ☐ **mismatched embedding algorithm**

real world

Motivation of Our Work

Steganalysis Really Win the battle?

Train	Test	JRM [1]	PF-274 [2]
Source	Source	100%	98.5%
Target	Target	99.5%	97.3%
Source	**Target**	**73.5%**	**70.5%**

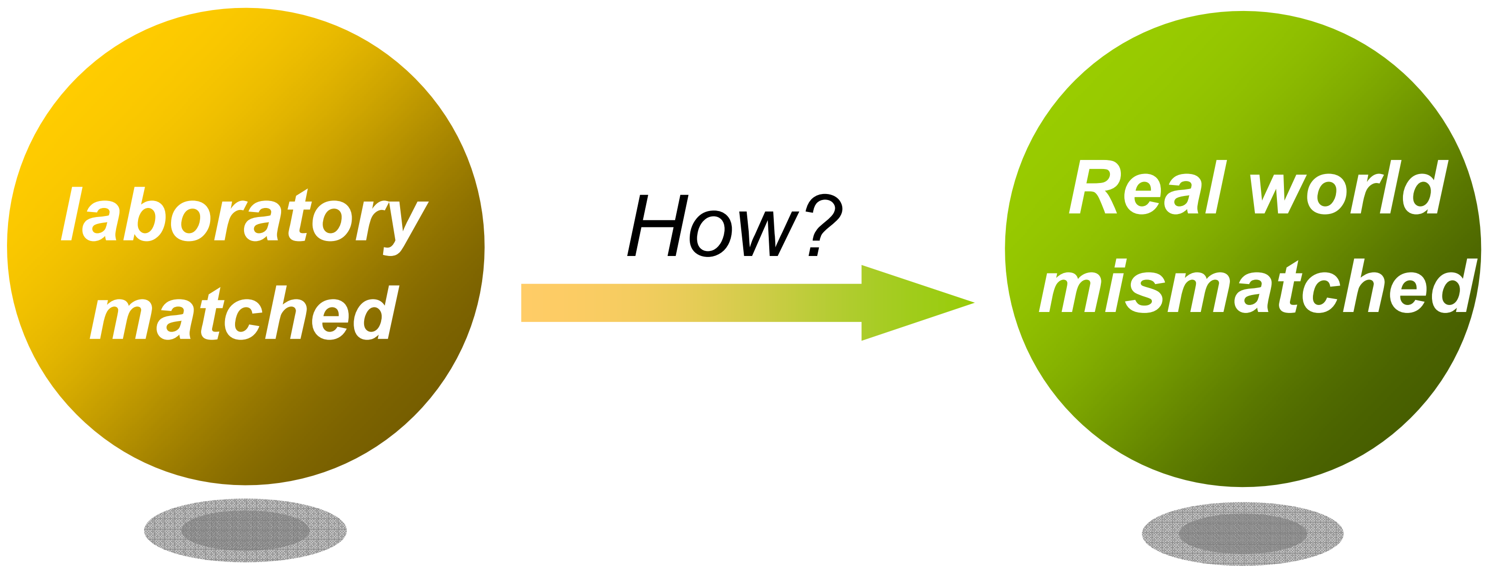
•[1] [Kodovsky and Fridrich 12]:Steganalysis of jpeg images using rich models
•[2] [Tomas Pevny and Jessica Fridrich 07]:Merging markovand dct features for multi-class jpeg steganalysis

Motivation of Our Work

New Challenge in Steganalysis

State-of-the-art steganalysis method could not be used effectively in the real world. Moving steganalysis from the laboratory to the real world.

[Andrew D. Ker, Patrick Bas, Rainer Bohme, Remi Cogranne, Scott Craver Tomas Filler, Jessica Fridrich, Tomas Pevny 13]: Moving steganography And steganalysis from the laboratory to the real world.



Motivation of Our Work

Related Work

[Ivans Lubenko, Andrew D. Ker 13]: Steganalysis with Mismatched Covers: Do Simple Classifiers Help?

Large data

- large samples
- diverse enough
- rich features

+

Simple classifier

- Ensemble Fisher Linear Discriminant
- Online Ensemble Average Perceptron

→ general model

Limitation: It costs much labor to collect images for such a training set.

Can we train a model robust to mismatched steganalysis using a small set of samples? ——only a single set, not diverse, small number.

Contents

Proposed Method

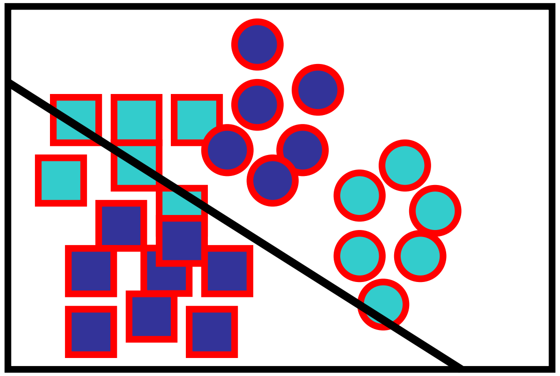
Main idea

- Train set: Source domain

$$\mathcal{D}_S = \{(x_i, y_i), i = 1, 2, \cdots, N\} \sim P_S(X, Y)$$

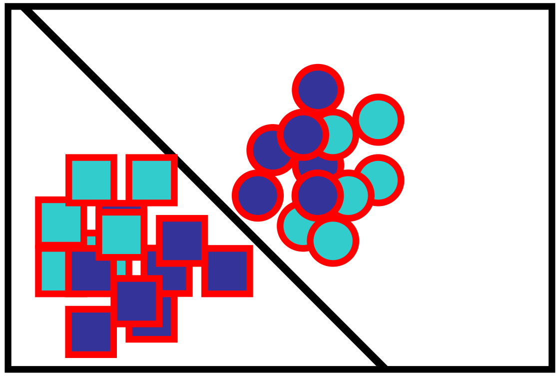
- Test set: Target domain

$$\mathcal{D}_T = \{(x_i, ?), i = 1, 2, \cdots, M\} \sim P_T(X, Y)$$



*The two distributions are **not** the same!*

?



The two distributions are similar!

Proposed Method

domain adaptation & transfer learning

- For Mismatch in other area:

Natural Language Processing
Ex. [J. Blitzer et al EMNLP 2006]

Object recognition
Ex. [R. Gopalan et al ICCV 2011]

Video analysis
Ex. [Jeff Donahue et al CVPR 2013]

Text classification
Ex. [Pan et al IEEE Tran-NN 2011]

domain adaptation & transfer learning

- Learning a shared representation

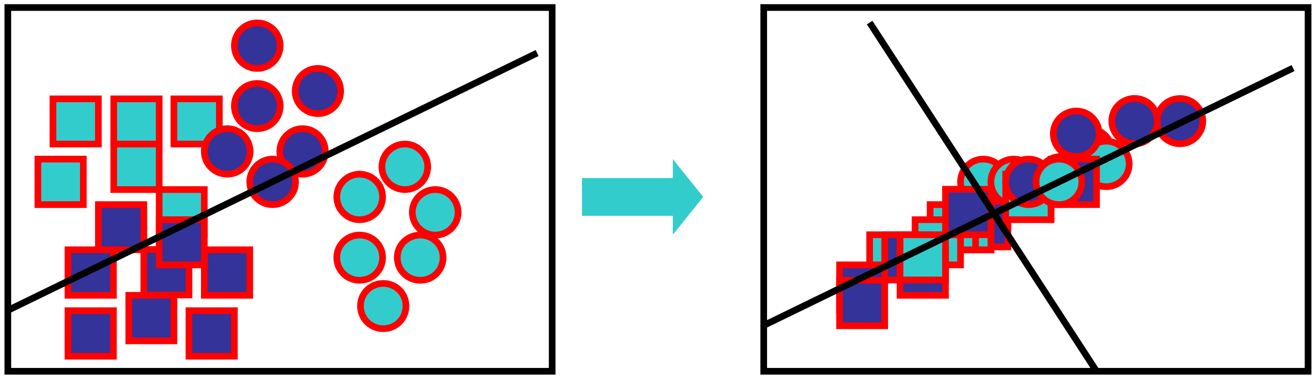
Assumption: a latent feature space exists in which classification hypotheses fit both domains.

$$\min_{f} \quad |P_S(f(X),Y)\text{-}P_T(f(X),Y)|$$

Proposed Method

Main idea

- Such a latent feature space leads to loss of some information, and may not be sensitive to embedding.

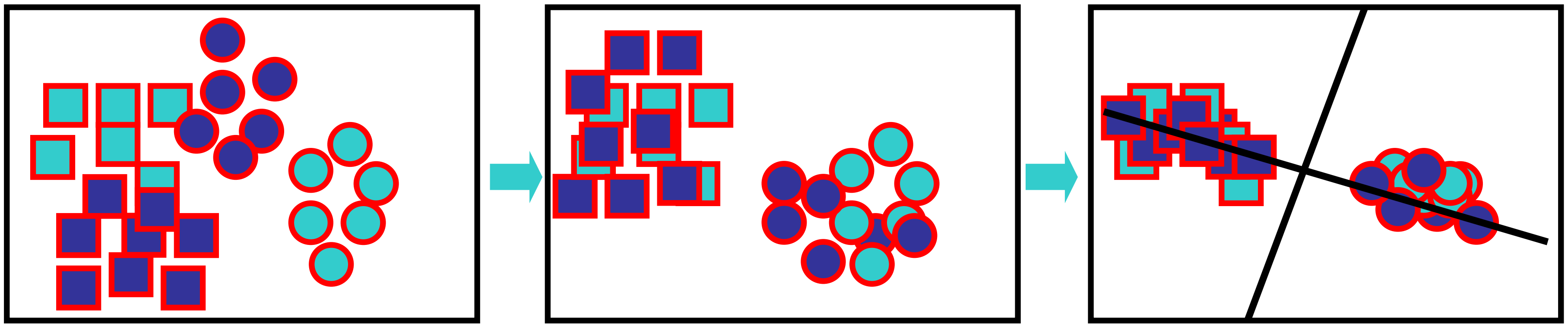


- According to target domain, transform source domain to an intermediate domain.
- Then find a latent feature space between target domain and intermediate domain.

Proposed Method

Generalized Transfer Component Analysis

- Domain Alignment: transform source domain to intermediate domain.
- Learn Shared Feature Space: find a latent feature space between target and intermediate domain.
- Map Samples into the Feature Space
- Construct Classifier and Make Decision in the New Feature Space



Proposed Method

Domain Alignment

- The aim is to transform source to an intermediate that is close to target.

•similar to 0-1 normalization liner transformation to hold the feature sensitivity to different categories.

•Objective :

$$\boxed{\begin{aligned} E(\varphi(X_s),Y) &= E(X_t,Y) \\ \sigma(\varphi(X_s),Y) &= \sigma(X_t,Y) \end{aligned}}$$

$$\varphi(x_s^i) = (x_s^i - E(X_s,y_i))\frac{\boxed{\sigma(X_t,y_i)}}{\sigma(X_s,y_i)} + \boxed{E(X_t,y_i)}$$

No labels in test set (target domain). We can't get the values.

Proposed Method

Domain Alignment

- Objective :

$$E_s(\psi(X_s)) = E(X_t)$$

$$\sigma(\psi(X_s)) = \sigma(X_t)$$

- Liner transformation :

$$\varphi(x_s^i) = (x_s^i - E(X_s))\frac{\sigma(X_t)}{\sigma(X_s)} + E(X_t)$$

Train Model

$$p_{s\to t}(y \mid x_t)$$

$$E(X_t \mid Y) \approx \frac{1}{\sum p(y \mid x_t^i)} \sum x_t^i p(y \mid x_t^i)$$

$$\sigma(X_t \mid Y) \approx \sqrt{\frac{1}{\sum p(y \mid x_t^i)} \sum (x_t^i \text{-} E(X_t, Y))^2 p(y \mid x_t^i)}$$

Proposed Method

Find shared feature space

- Objective:

$$\min_{f} \quad |P_S(f(X),Y)\text{-}P_T(f(X),Y)|$$

- Simplify:

$$\boxed{\begin{aligned} P(X,Y) &= P(Y\,|\,X)P(X) \\ P_s(Y\,|\,X) &= P_t(Y\,|\,X) \end{aligned}} \Rightarrow \min_{f} \quad |P_S(f(X))\text{-}P_T(f(X))|$$

- Measure the distance of two distribution:

$$\mathrm{Dis}(P_S(X),P_T(X)) = \|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(x_s^i) - \frac{1}{n_t}\sum_{i=1}^{n_t}\phi(x_t^i)\| \qquad \phi(.) \rightarrow \textit{RHKS}$$

$$Dis(P_S(X),P_T(X)) = \mathrm{trace}(KL)$$

K. M. Borgwardt, A. Gretton, M. J. Rasch, H. P. Kriegel, B. Scholkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," Bioinformatics, 2006

Proposed Method

Find shared feature space

- define a non-liner kernel feature extraction matrix *W* as transformation:

$$X_{new} = KW$$

- Update the new *K*:

$$K_{new} = X_{new} X_{new}^T = KWW^T K$$

- Update the new *distance*:

$$\mathrm{Dis}(P_S(X), P_T(X)) = trace(KL) = trace(KWW^T KL)$$

$$\Rightarrow \quad \min_W \quad trace(KWW^T KL)$$

S. Pan, I. Tsang, J. Kwok, and Q.Yang, “Domain adap-tation via transfer component analysis,” IEEE Trans-actions on Neural Networks, 2011

Proposed Method

Find shared feature space

- To avoid the solution *W=0,* we add a constrain that which can preserve (or maximize) the initial data variance in the new space:

$$W^T KHKW = I$$

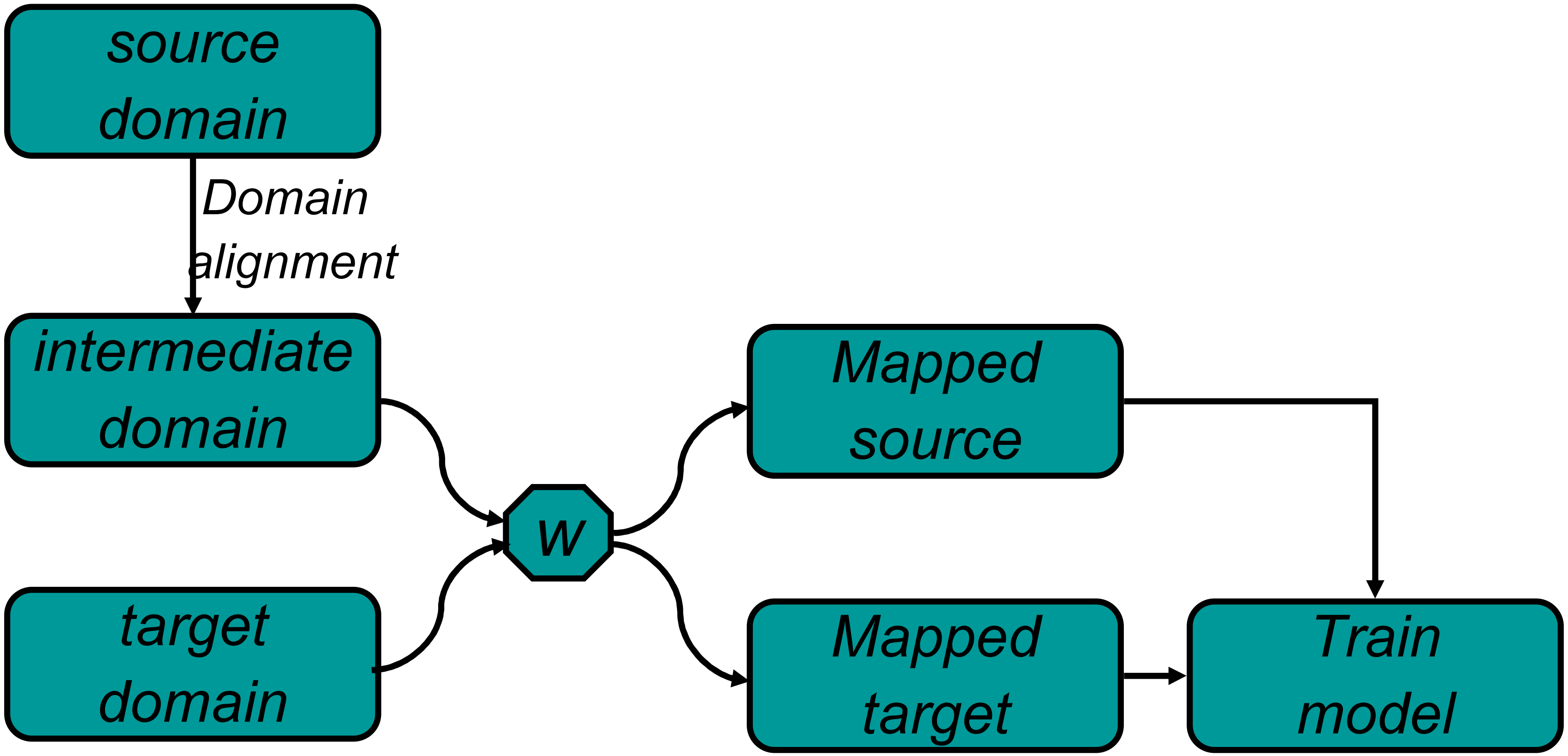
- The final kernel learning problem is then set up as:

$$\min_{W} \quad tr(W^T W) + \mu tr(KWW^T KL)$$
$$s.t. \qquad W^T KHKW = I$$

$\Rightarrow$ $W \to (I + \mu KLK)^{-1} KHK$ (M leading eigenvectors)

S. Pan, I. Tsang, J. Kwok, and Q.Yang, “Domain adap-tation via transfer component analysis,” IEEE Trans-actions on Neural Networks, 2011

Proposed Method



Contents

Experiments and Results

Experimental Setup

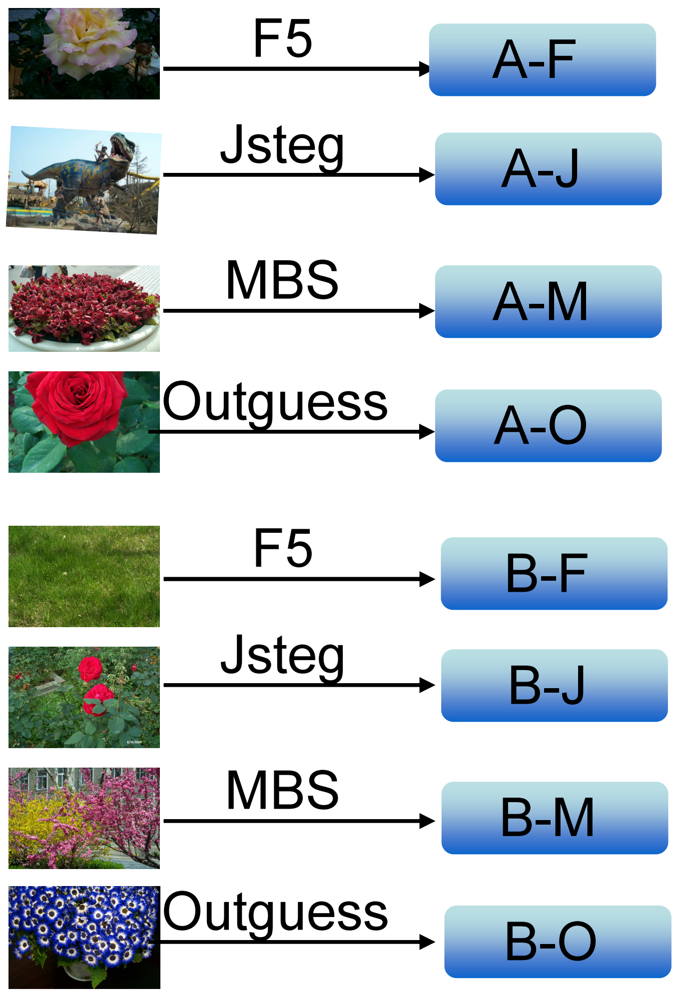
- Database: eight mismatched domains

1	1	1	1	1	2	3	3
1	1	1	1	1	3	3	3
1	1	1	1	1	3	3	3
1	1	1	1	2	4	4	3
1	1	3	4	4	6	6	4
1	2	3	3	4	5	6	5
2	3	4	4	5	6	6	5
5	5	5	5	5	5	5	5

Set A: 1800

9	6	5	9	13	22	29	35
6	6	8	11	15	33	34	30
8	7	9	13	22	33	39	31
8	9	12	16	28	49	45	34
10	12	21	32	39	61	58	42
13	19	31	36	45	58	63	51
28	36	44	49	58	68	66	55
41	52	54	55	62	56	57	54

Set B:1800



Each domain
Cover:450
Stego:450

Experiments and Results

Experimental Setup

- Database: eight mismatched domains
- Features: PF274 + our approach (GTCA)
- Classifier: lib-SVM
- Approach compared with:
  - Orig-Fea: PF-274+ lib-SVM
    - *[Pevny and Fridrich 07]:Merging markov and dct features for multi-class jpeg steganalysis*
  - OEAP: JRM features + OEAP
    - *[Kodovsky and Fridrich 12]:Steganalysis of jpeg images using rich models*
    - *[Ivans Lubenko, Andrew D. Ker 13]: Steganalysis with Mismatched Covers: Do Simple Classifiers Help?*
  - TCA: PF274+ TCA+ lib-SVM
    - *[Pan et al 2011] Domain adap-tation via transfer component analysis*

Experiments and Results

Mismatched Experiment 1

- Mismatched covers: different quantization table

Train→Test	A-F→B-F	A-J→B-J	A-M→B-M	A-O→B-O
Orig-Fea	0.505	0.515	0.515	0.505
OEAP [4]	0.500	0.515	0.523	0.515
TCA [17]	0.500	0.505	0.549	0.827
GTCA	0.884	0.965	0.931	0.944

Train→Test	B-F→A-F	B-J→A-J	B-M→A-M	B-O→A-O
Orig-Fea	0.505	0.545	0.535	0.515
OEAP [4]	0.525	0.505	0.505	0.545
TCA [17]	0.525	0.825	0.899	0.515
GTCA	0.787	0.975	0.951	0.865

Experiments and Results

Mismatched Experiment 2

- Mismatched stegos: different embedding algorithm

Train→Test	B-F→B-M	B-F→B-O	B-J→B-M	B-J→B-O
Orig-Fea	0.695	0.705	0.533	0.515
OEAP [4]	0.833	0.755	0.553	0.535
TCA [17]	0.865	0.785	0.655	0.602
GTCA	0.885	0.875	0.835	0.845

Train→Test	B-J→B-F	B-M→B-F	B-M→B-O	B-O→B-F
Orig-Fea	0.515	0.635	0.870	0.541
OEAP [4]	0.572	0.653	0.905	0.835
TCA [17]	0.755	0.845	0.855	0.775
GTCA	0.745	0.885	0.983	0.877

Experiments and Results

Mismatched Experiment 3

- Mismatched covers and stegos: different quantization table and different embedding algorithm

Train→Test	A-F→B-M	A-F→B-O	A-J→B-M	A-J→B-O
Orig-Fea	0.495	0.510	0.500	0.500
OEAP [4]	0.535	0.545	0.523	0.515
TCA [17]	0.505	0.500	0.500	0.502
GTCA	0.805	0.837	0.785	0.753

Train→Test	A-J→B-F	A-M→B-F	A-M→B-O	A-O→B-F
Orig-Fea	0.500	0.510	0.535	0.512
OEAP [4]	0.523	0.515	0.550	0.515
TCA [17]	0.504	0.530	0.559	0.575
GTCA	0.733	0.735	0.922	0.807

Contents

Summary

- Mismatched steganalysis
  - Important in real application
  - Traditional steganalysis methods perform badly
  - Two distributions are not the same
- Generalized Transfer Component Analysis (GTCA)
  - Learn new representations to correct mismatches
  - A small set of training samples
  - Empirically successful
- New Strategy for Mismatched Steganalysis
  - Domain adaptation, transfer learning

Thank you!