SECURE STEGANOGRAPHY: STATISTICAL RESTORATION OF THE SECOND ORDER DEPENDENCIES FOR IMPROVED SECURITY

A. Sarkar, K. Solanki, U. Madhow, S. Chandrasekaran and B. S. Manjunath

Department of Electrical and Computer Engineering,
University of California at Santa Barbara
Santa Barbara, CA 93106

ABSTRACT

We present practical approaches for steganography that can provide improved security by closely matching the second-order statistics of the host rather than just the marginal distribution. The methods are based on the framework of *statistical restoration*, wherein a fraction of the host symbols available for hiding is actually used to restore the statistics; thus reducing the rate, but providing security against steganalysis. We establish correspondence between steganography and the earth-mover's distance (EMD), a popular distance metric used in computer vision applications. The EMD framework can be used to define the optimum flow (modifications) of the host symbols for compensation. This formulation is used for image steganography by restoring the second-order statistics of the block-wise discrete cosine transform (DCT) coefficients. Some practical limitations of this approach (such as computational complexity and difficulty in dealing with overlapping coefficient pairs) are noted, and a new method is proposed that alleviates these deficiencies by identifying the coefficients to modify based on a local compensation criterion. Experimental results on several thousand natural images demonstrate the utility of the presented methods.

***Index Terms*—** earth mover's distance, joint compensation, statistical restoration, steganalysis, steganography.

1. INTRODUCTION

We consider the problem of secure communication or *steganography*, in which a secret message is hidden into an innocuous looking *host* or *cover* to get a *composite* or *stego* image such that the very presence of communication is not revealed. To detect the presence of embedded data, *steganalysis* techniques exploit the changes in the host statistics due to hiding. In this paper, we present practical techniques for image steganography that provide improved security in comparison to many prior schemes by closely matching the second-order distribution rather than just the marginal statistics. The techniques are based on a general steganography framework called statistical restoration, initially proposed in our prior publications ([1, 2, 3]) and applied to marginal statistics .

There have been several approaches in the past that attempt to restore the first-order statistics so as to resist histogram-based steganalysis. These include Provos' OutGuess algorithm [4], Eggers et al's histogram-preserving data-mapping [5], Franz's suggestion [6] of hiding in independent pairs of values, Guillon's idea [7] of companding to a uniform distribution prior to quantization-based hiding, and Wang and Moulin's stochastic quantization index modulation (QIM) [8]. Some of these approaches are limited by their inability to handle continuous host data, while others cannot achieve exact host probability mass function (PMF) when communicating at high rates. Many of these schemes are also fragile against any noise or attacks. In our prior work [1, 2, 3], a statistical restoration scheme was proposed that addresses some of these deficiencies. It was shown that one could achieve zero Kullback-Leibler (K-L) divergence between the host and the stego PMFs while hiding at high rates. The scheme was also robust against distortion-constrained attacks.

In this paper, we establish correspondence between the statistical restoration framework and the earth-mover's distance (EMD) [9], a distance metric widely used for image similarity search. This allows us to leverage the analytical results derived for EMD over the past years. We present a practical method for image steganography based on the EMD formulation for restoring two-dimensional statistics of the block-wise discrete cosine transform (DCT) coefficients. This is aimed at defeating steganalysis methods that use statistics derived from pairs of DCT coefficients as features [10]. There are, however, some practical issues in applying the optimal EMD *flow* (i.e. optimal modifications) for DCT coefficients. These include high computational complexity and difficulty in dealing with overlapping coefficient pairs. To deal with these limitations, we propose a new approach that modifies the compensation coefficients based on a local criterion. This approach is sub-optimal (in terms of flow computation), but is computationally tractable, and can closely match the distribution of overlapping pairs. This method also restores intra and inter-block based correlation histograms (hence termed *joint* compensation scheme), shown to be useful for steganalysis in [10].

We test our steganographic methods on a set of 4500 images, using supervised learning to train classifiers for steganalysis. We report results for several different classifiers trained on different features, such as 1-D PMFs, overlapping and non-overlapping pair based 2-D PMFs, and PMFs derived from joint (intra and inter-block based) correlations. It is seen that both the EMD-flow based and the joint compensation based schemes are quite effective in evading steganalysis based on these features.

2. PROBLEM FORMULATION

In the statistical restoration framework, the set of host symbols X is divided into two disjoint sets: H for hiding and C for compensation. Data is embedded using the hiding function f_1 into the hiding set H to get $\hat{H}$, as shown in (1) later. We divide the host symbol set X into 2-D bins and find their respective bin-counts (number of terms per bin). We use $B_X(i,j)$ to denote the bin-count of the $(i,j)^{th}$ bin of X. Since the normalized bin-count gives the PMF, compensating for the bin-counts is equivalent to restoring the PMF. The aim is to find

This research is supported in part by a grant from ONR # N00014-05-1-0816.

the function f_2 that modifies C to $\hat{C}$ such that the 2-D PMF $P_Y(Y)$ (Y having been defined in (1)), obtained after hiding and compensation, is same as that of X. To evade second order steganalysis, a lower hiding rate is employed compared to the 1-D case.

$$X = H \cup C,\ \hat{H} = f_1(H),\ \hat{C} = f_2(C),\ Y = \hat{H} \cup \hat{C} \quad (1)$$

$$H \cap C = \phi \Rightarrow B_X(i,j) = B_H(i,j) + B_C(i,j),\ \forall(i,j) \quad (2)$$

$$\hat{H} \cap \hat{C} = \phi \Rightarrow B_Y(i,j) = B_{\hat{H}}(i,j) + B_{\hat{C}}(i,j),\ \forall(i,j) \quad (3)$$

$$\text{To obtain } P_Y = P_X, \text{ we need } B_Y(i,j) = B_X(i,j),\ \forall(i,j) \quad (4)$$

$$\Rightarrow B_{\hat{C}}(i,j) = B_C(i,j) + B_H(i,j) - B_{\hat{H}}(i,j),\ \forall(i,j) \quad (5)$$

There are two main constraints that must be satisfied by the hiding and compensation functions (f_1 and f_2, respectively):

- *Perceptual constraint*: The embedding process must induce minimal perceptual distortion to the host signal, which limits the maximum average (and/or peak) distortion that the functions f_1 and f_2 may incur.
- *Statistical constraint*: The divergence introduced by the embedding process must be less than a small number ϵ (proposed in [11]). This constraint defines the goal of the function f_2.

3. STATISTICAL RESTORATION AND THE EARTH MOVER'S DISTANCE

The EMD [9] between two PMF's is defined as the minimum "work" done in converting one PMF to the other. Here, *work* refers to the redistribution of weights among the various bins in the discrete distribution. The solution to the EMD problem returns the optimal transportation flows among the bins. For our statistical restoration problem, we have to convert a 2-D histogram B_C to $B_{\hat{C}}$, according to (5), the normalized histogram being the PMF. Thus, EMD provides the optimum way of redistributing weights in B_C to obtain $B_{\hat{C}}$.

Let S and T denote two 2-D signatures, each having M clusters. The weight of each cluster is the fraction of points it contains. Let the center for the k^{th} ($k = (i,j)$) cluster of S be $\{s_i, s_j\}$ while the ℓ^{th} ($\ell = (m,n)$) cluster center of T is denoted by $\{t_m, t_n\}$. The square Euclidean distance between the k^{th} cluster center of S and the ℓ^{th} cluster center of T is called $d_{k\ell}$.

$$d_{k\ell} = (s_i - t_m)^2 + (s_j - t_n)^2,\ k = (i,j),\ \ell = (m,n) \quad (6)$$

The EMD problem is "optimally" changing S (considered as the source) to make it as similar as possible to T (the target). For our problem, the source S is the PMF P_C of the compensation coefficients while the target T is $P_{\hat{C}}$, the PMF of $\hat{C}$. The weight of each bin is the PMF value for that bin. Our aim is to find a flow matrix $F = [f_{k\ell}]$, where $f_{k\ell}$ is the flow from the k^{th} bin of S to the ℓ^{th} bin of T that minimizes the total *work* done:

$$WORK(S,T,F) = \sum_{k=1}^{M} \sum_{\ell=1}^{M} d_{k\ell} f_{k\ell} \quad (7)$$

Thus, EMD gives precisely the optimum flows from the bins of C to $\hat{C}$ that match P_X to P_Y, where X and Y are defined in (1), under the minimum mean-squared error (MMSE) criterion. The above formulation can be generalized to the n-D case.

3.1. EMD Formulation for Image Steganography

The EMD-based statistical compensation is used to restore higher order statistics for image steganography. We compute the DCT of 8×8 blocks in the image, divide by a certain quality factor matrix, round off the terms and hide data using odd-even embedding, as in [1]. The aim is to compensate for the intra-block based 2-D statistics in the quantized DCT domain. We discuss some important issues regarding the implementation of 2-D EMD based statistical restoration.

• *Overlapping Pairs:* For accurate 2-D PMF estimation, we need to consider all possible consecutive pairs. However, if we take overlapping pairs, then, when considering a flow from one bin to another, the bin-counts in adjacent bins are inadvertently modified. The remedy is to consider non-overlapping pairs of DCT coefficients per block, as shown in Fig. 1.

• *Computational Complexity:* Let the number of bins be N in the 1-D case. By considering all possible pairs in the 2-D case, we end up with N^2 bins. Due to the increased number of bins, storing the cost transportation matrix ($N^2 \times N^2$) becomes difficult and solving the 2-D EMD problem becomes computationally intensive. We have used 350 bins in our 2-D implementation [1] due to the memory constraints. We consider the top 350 bins for which the bin-count difference between the source and target PMFs is most significant. Thus, we can only obtain an approximate 2-D PMF-matching. Also, if the first order statistics are not fully restored and the mismatch is high enough, detection can be achieved using only 1-D PMFs. Thus, we need to solve the reduced second-order PMF matching problem, under the constraint that first order PMFs are fully matched (or closely enough to resist steganalysis).

• *Perceptual limitations:* For the 2-D PMF case, during EMD-based compensation, perturbations of magnitude greater than 2 may occur for the quantized DCT coefficients. For perceptual transparency, we limit the perturbations to ± 1.

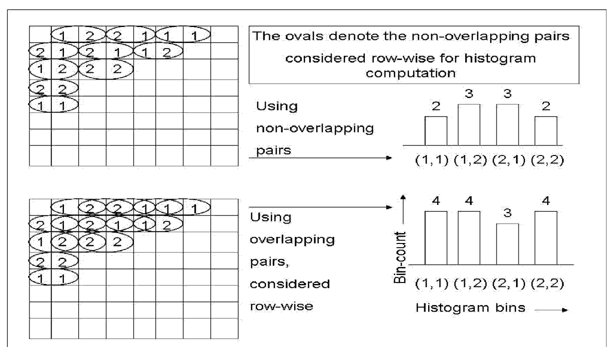

Fig. 1. Histogram computation row-wise, considering overlapping and non-overlapping pairs of coefficients, per 8×8 blocks

4. JOINT INTRA- AND INTER-BLOCK COMPENSATION

In the EMD-based method, we considered the row-wise pairing among successive DCT coefficients in the same block such that the intra-block dependency was accounted for. In [10], scanning techniques have been discussed to construct a matrix of DCT coefficients which captures both intra and inter-block correlations.

- The AC DCT coefficients obtained from a zigzag scan along the 8×8 block are generally in descending order of magnitude. Hence, the (intra-block) correlation among consecutive terms will be more significant if the coefficients are arranged in the zigzag scan order.

[1]The EMD implementation used is available online at http://ai.stanford.edu/∼rubner/emd/default.htm

- In an image, two neighboring blocks are very likely to be similar due to the high low-frequency content. Therefore, we can expect high (inter-block) correlation between the same AC DCT term among neighboring blocks. An alternate block scanning method is shown in Fig. 2 which ensures that sequentially scanned blocks are spatially correlated.

As proposed in [10], we arrange the DCT coefficients in a $N_r \times N_c$ matrix A (shown in Fig. 4). The first N_c AC DCT terms occurring in the zigzag scan order in the same block of an image are placed in the same row. The N_r 8×8 blocks in the image scanned in the alternate sequence constitute the rows. Thus, two consecutive terms along the same row (column) provide the intra(inter)-block correlation. The row-wise and column-wise PMFs computed using overlapping pairs of elements of the matrix A have been shown to be useful for steganalysis [10]. If we consider non-overlapping pairs either along the rows or columns of A, intra-block or inter-block compensation can be done respectively using the EMD formulation, but not both simultaneously.

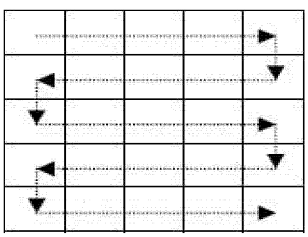

Fig. 2. Alternate scanning along the rows of an image to obtain spatially correlated blocks [10]

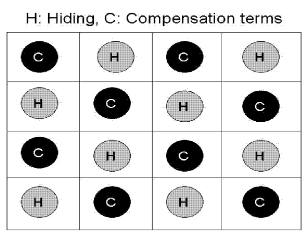

Fig. 3. The intra and inter-block based matrix A has H and C terms arranged alternately

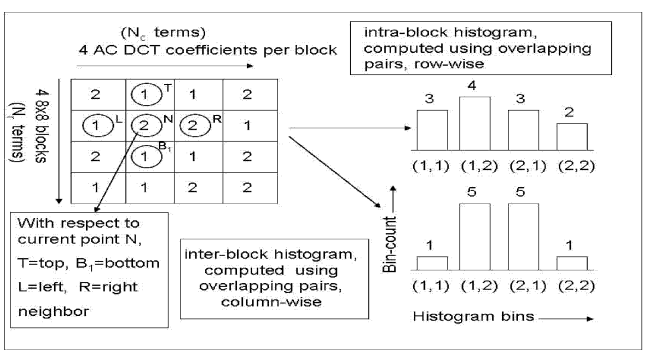

Fig. 4. Explanation of the intra-block and inter-block histogram computation for the $N_r \times N_c$ matrix A; in this example, N_r=N_c=4

For allowing overlapping pairs in the PMF computation, we divide the matrix A into two classes 'C' and 'H', as in Fig. 3. If a certain point belongs to 'C', its 4 nearest (D_4) neighbors all belong to 'H' while the 4 diagonal neighbors belong to 'C' and vice versa. We perform hiding in certain terms in 'H' while only the elements in 'C' are used for compensation. When we modify a certain term in 'C', it affects the bin-count in pairs comprising itself and its D_4 neighbors. At each point, considering the bin-count difference, computed between original and target 2-D PMFs, for these 4 pairs, we decide on whether to perturb a certain value by $\pm$ 1 (perturbation is limited to $\pm$ 1 for perceptual transparency) or retain it. Thus, the decision taken at each compensation point is optimal provided that none of its D_4 neighbors change, which we ensure by constraining its D_4 neighbors to belong to the non-compensation stream.

Let B_{intra} and B_{inter} denote the intra-block and inter-block bin-counts obtained using the matrix A.
$B_{intra}(a,b) = \sum_{i,j} \mathcal{I}_{ij},\ \mathcal{I}_{ij} = 1$ if $\{A_{ij} = a, A_{i,j+1} = b\}$ else 0
$B_{inter}(a,b) = \sum_{i,j} \mathcal{J}_{ij},\ \mathcal{J}_{ij} = 1$ if $\{A_{ij} = a, A_{i+1,j} = b\}$ else 0
where A_{ij} is the element in the i^{th} row and j^{th} column of A. After data hiding without compensation, let the modified intra-block and inter-block bin-counts of the matrix A be called B'_{intra} and B'_{inter}, respectively. Let us consider the element A_{ij}, which equals N, while (as in Fig. 4) its D_4 neighbors are $A_{i,j-1} = L$ (left), $A_{i,j+1} = R$ (right), $A_{i-1,j} = T$ (top) and $A_{i+1,j} = B_1$ (bottom).

Since we allow a perturbation of only ± 1, N can be mapped to one of $\{N-1, N, N+1\}$. We compute the 4 bin-count difference values for $N' \in \{N-1, N, N+1\}$:

$$\mathcal{D}(N',1) = B_{intra}(L,N') - B'_{intra}(L,N')$$
$$\mathcal{D}(N',2) = B_{intra}(N',R) - B'_{intra}(N',R)$$
$$\mathcal{D}(N',3) = B_{inter}(T,N') - B'_{inter}(T,N')$$
$$\mathcal{D}(N',4) = B_{inter}(N',B_1) - B'_{inter}(N',B_1)$$

If the point N lies on the boundary, and lacks one or more of the D_4 neighbors, we just replace the corresponding $\mathcal{D}$ term with 0.

The squared difference between the original and modified histograms for the intra and inter-block cases are considered. For every modification of N to one of $\{N-1, N, N+1\}$, there are 4 (2 intra, for L and R, and 2 inter, for T and B_1) histogram entries that are changed. So, a maximum of 4×3=12 $\mathcal{D}$ terms may vary depending on how N is changed, as in the expression for the squared error cost function J in (8). When N is changed to $(N\pm1)$, the B'_{intra} and B'_{inter} terms, associated with $(N\pm1)$ and its D_4 neighbors, are increased by 1 and the corresponding bin-counts, associated with N and its D_4 neighbors, are decreased by 1 - this explains the use of the $I_{\delta,k}$ indicator function in (8). N is converted to that N_{opt} (9) for which the squared difference term J is minimized.

$$J(N+\delta) = \sum_{i=1}^{4} \{\mathcal{D}(N,i) + 1 - I_{\delta,0}\}^2 + \sum_{i=1}^{4} \{\mathcal{D}(N-1,i) - I_{\delta,-1}\}^2 + \sum_{i=1}^{4} \{\mathcal{D}(N+1,i) - I_{\delta,1}\}^2,\ \delta = \{-1,0,1\} \tag{8}$$

where the indicator function $I_{\delta,k} = 1$ if $\delta = k$ and $= 0$ otherwise

$$N_{opt} = \underset{N',\ N' \in \{N-1,N,N+1\}}{\arg\min} J(N') \tag{9}$$

We repeat this process to obtain a locally optimal solution for each compensation location of A. The collection of all these locally optimal solutions provides an answer to the following problems:

- 2-D histogram compensation considering overlapping pairs,
- simultaneous intra-block and inter-block compensation and
- deciding which DCT coefficients should be perturbed. EMD just provides the flow from one bin to the other but does not suggest the particular elements to modify.

5. EXPERIMENTS AND RESULTS

For evaluation of our steganographic schemes, we use support vector machine (SVM) based steganalysis to detect the stego images. The stego images use a hiding fraction of 10% and have been statistically compensated using three different schemes (Table 1). We use 4500 images for our experiments - half for training and the other half for testing. Both the training and testing sets have half the images as cover and the other half as stego. During the training phase, we develop separate SVM classifiers trained on each feature used for steganalysis in Table 1. The SVM classifiers are then used to distinguish between cover and stego images in the testing phase.

While computing 2-D histograms for DCT coefficients, we only consider those with magnitude less than T (threshold T=30 is used). Since the distribution of the DCT coefficients falls off sharply for higher values, higher valued terms may be ignored in PMF estimation. In EMD-based 2-D compensation, we consider 20 AC DCT terms per block (Fig. 1). The top 350 bins, for which the bin-count difference between the source and target PMF's is maximum (Sec. 3.1), may vary from image to image. Due to the high low-frequency content in an image, the $(0,0)^{th}$ bin and most of the bins near it will have a large bin-count. Hence, we consider a square window of bins, (with $(0,0)$ as the center and each side of length 21) consisting of bin-counts of the bins it contains as the feature vector.

For the joint intra-inter based compensation problem, we take the first 15 AC DCT coefficients that occur during zigzag scan per 8×8 block. We then consider square windows of size 13×13 and centered at the $(0,0)^{th}$ bin for both the intra and inter-block histogram matrices. This results in a feature vector of length $2\times13^2 = 338$ used for joint compensation.

In Table 1, "Joint", "EMD" and "1D" refer to the intra and inter block PMF restoration, 2-D non-overlapping pair based intra-block PMF restoration and 1-D PMF restoration based steganography methods, respectively. We test the methods against SVM classifiers trained on the following features: Intra(Inter) - overlapping pair based intra(inter)-block histogram; Joint - accounts for both intra and inter-block histograms; 1D - first order PMF (256 bins), using first 15 AC DCT coefficients per block (zigzag scan order) with values in [-128,127]; 2D - non-overlapping pair based intra-block histogram.

Table 1. Comparison of the performance of the three compensation methods, when steganalysis experiments are performed using five different features. P_{FA} and P_{miss} represent the probability of false alarm and of missed detection, respectively. For undetectable hiding, the detector will be reduced to random guessing and the total detection error, $(P_{FA} + P_{miss})$, will be close to 1.The rows contain the five features while the columns denote the three methods, which are not to be confused though they have some names in common.

Feature	P_{FA}			P_{miss}			$(P_{FA} + P_{miss})$	
Used	Joint	EMD	1D	Joint	EMD	1D	Joint	EMD
Intra	0.21	0.04	0.04	0.59	0.28	0.31	0.80	0.32
Inter	0.27	0.14	0.16	0.58	0.32	0.32	0.85	0.46
Joint	0.28	0.04	0.04	0.52	0.27	0.30	0.80	0.31
1D	0.37	0.33	0.28	0.58	0.65	0.72	0.95	0.98
2D	0.24	0.35	0.09	0.19	0.47	0.28	0.43	0.82

From Fig. 5, we observe that for steganalysis based on intra-block, inter-block and joint matrices (overlapping pair based) features, the joint compensation based steganography scheme performs better than the EMD-based scheme as we explicitly compensate for these features. For 1-D PMF as input, both the schemes do well, while for non-overlapping pair based 2-D intra-block PMF (feature described in Fig. 1), the EMD scheme does better.

We end this section with a brief note on the relative computational costs of the two methods. For a n bin EMD problem, the complexity $\mathcal{C}_{EMD}$ is at best $O(n^3 \log n)$ [9]; for joint compensation, the complexity $\mathcal{C}_{joint}$=$O(|A|)$, $|A|$ being the cardinality of the joint correlation based matrix A (explained in Sec. 4). E.g. let the image size be 512×512, with 15 AC DCT terms being considered per block, while there are 350 bins in the EMD formulation. Then, $\frac{\mathcal{C}_{EMD}}{\mathcal{C}_{joint}} \approx 10^4$.

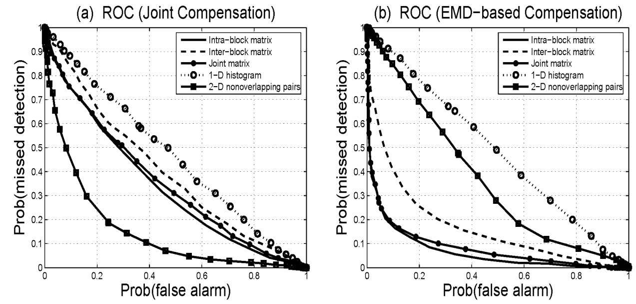

Fig. 5. Comparison of detection curves for a variety of features using (a) joint compensation and (b) EMD-based compensation methods

Also, the complexity for the joint compensation scheme increases linearly with the image size due to the dependence on $|A|$.

6. CONCLUSIONS

In this paper, we demonstrate practical steganographic methods that provide improved security by closely matching the second order statistics. We establish correspondence between EMD and steganography, an important contribution of this paper. This opens the doors for using the known results on EMD for steganography. An avenue of future work is to exploit the bounds available for EMD to get more computationally tractable solutions for steganography .

References

[1] K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Statistical restoration for robust and secure steganography," in *Proc. ICIP*, 2005, pp. II–1118–21.

[2] K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Probably secure steganography: Achieving zero K-L divergence using statistical restoration," in *Proc. ICIP*, 2006, pp. 125–128.

[3] K. Sullivan, K. Solanki, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Determing achievable rates for secure, zero-divergence, steganography," in *Proc. ICIP*, 2006, pp. 121–124.

[4] N. Provos, "Defending against statistical steganalysis," in *10th USENIX Security Symposium*, Washington DC, USA, 2001.

[5] J. J. Eggers, R. Bauml, and B. Girod, "A communications approach to image steganography," in *Proc. of SPIE*, San Jose, CA, 2002.

[6] E. Franz, "Steganography preserving statistical properties," in *5th International Working Conference on Communication and Multimedia Security*, 2002.

[7] P. Guillon, T. Furon, and P. Duhamel, "Applied public-key steganography," in *Proc. of SPIE*, San Jose, CA, 2002.

[8] Y. Wang and P. Moulin, "Steganalysis of block-structured stegotext," in *Proc. of SPIE*, San Jose, CA, 2004.

[9] Y. Rubner, C. Tomassi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[10] D. Fu, Y. Q. Shi, D. Zou, and G. Xuan, "JPEG steganalysis using empirical transition matrix in block dct domain," in *International Workshop on Multimedia Signal Processing*, Victoria, BC, Canada, Oct. 2006.

[11] C. Cachin, "An information theoretic model for steganography," *LNCS: 2nd Int'l Workshop on Info. Hiding*, vol. 1525, pp. 306–318, 1998.