Steganalysis Using Image Quality Metrics

İsmail Avcıbaş, *Member, IEEE*, Nasir Memon, *Member, IEEE*, and Bülent Sankur, *Member, IEEE*

***Abstract*—We present techniques for steganalysis of images that have been potentially subjected to steganographic algorithms, both within the passive warden and active warden frameworks. Our hypothesis is that steganographic schemes leave statistical evidence that can be exploited for detection with the aid of image quality features and multivariate regression analysis. To this effect image quality metrics have been identified based on the analysis of variance (ANOVA) technique as feature sets to distinguish between cover-images and stego-images. The classifier between cover and stego-images is built using multivariate regression on the selected quality metrics and is trained based on an estimate of the original image. Simulation results with the chosen feature set and well-known watermarking and steganographic techniques indicate that our approach is able with reasonable accuracy to distinguish between cover and stego images.**

***Index Terms*—Analysis of variance, image quality measures, multivariate regression analysis, steganalysis, steganography, watermarking.**

I. Introduction

STEGANOGRAPHY refers to the science of "invisible" communication. Unlike cryptography, where the goal is to secure communications from an eavesdropper, steganographic techniques strive to hide the very presence of the message itself from an observer. Although steganography is an ancient subject, the modern formulation of it is often given in terms of the *prisoner's problem* [1] where Alice and Bob are two inmates who wish to communicate in order to hatch an escape plan. However, all communication between them is examined by the warden, Wendy, who will put them in solitary confinement at the slightest suspicion of covert communication. Specifically, in the general model for steganography, we have Alice wishing to send a *secret message* m to Bob. In order to do so, she "embeds" m into a *cover-object* c, to obtain the *stego-object* s. The stego-object s is then sent through the public channel.

The warden, Wendy, who is free to examine all messages exchanged between Alice and Bob, can be *passive* or *active*. A passive warden simply examines the message and tries to determine if it potentially contains a hidden message. If it appears that it does, she then takes appropriate action, else, she lets the message through without any action. An active warden, on the other hand, can alter messages deliberately, even though she may not see any trace of a hidden message, in order to foil any secret communication that can nevertheless be occurring between Alice and Bob. The amount of change the warden is allowed to make depends on the model being used and the cover-objects being employed. For example, with images, it would make sense that the warden is allowed to make changes as long as she does not alter significantly the subjective visual quality of a suspected stego-image.

It should be noted that the main goal of steganography is to communicate securely in a completely undetectable manner. That is, Wendy should not be able to distinguish in any sense between cover-objects (objects not containing any secret message) and stego-objects (objects containing a secret message). In this context, "*steganalysis*" refers to the body of techniques that are designed to distinguish between cover-objects and stego-objects. It should be noted that nothing might be gleaned about the contents of the secret message m. When the existence of hidden message is known, revealing its content is not always necessary. Just disabling and rendering it useless will defeat the very purpose of steganography. In this paper, we present a steganalysis technique for detecting *stego-images*, i.e., still images containing hidden messages, using image quality metrics. Although we focus on images, the general techniques we discuss would also be applicable to audio and video media.

Given the proliferation of digital images, and given the high degree of redundancy present in a digital representation of an image (despite compression), there has been an increased interest in using digital images as cover-objects for the purpose of steganography. The simplest of such techniques essentially embeds the message in a subset of the LSB (least significant bit) plane of the image, possibly after encryption [2]. It is well known that an image is generally not visually affected when its least significant bit plane is changed. Popular steganographic tools based on LSB like embedding vary in their approach for hiding information. For example *Steganos* and *Stools* use LSB embedding in the spatial domain, while *Jsteg* embeds in the frequency domain. Other more sophisticated techniques include the use of quantization and dithering. For a good survey of steganography techniques, the reader is referred to [2]. What is common to these techniques is that they assume a passive warden framework. That is they assume the warden Wendy will not alter the image. We collectively refer to these techniques as *passive warden steganography techniques*.

Conventional passive warden steganography techniques like LSB embedding are not useful in the presence of an active warden as the warden can simply randomize the LSB plane to thwart communication. In order to deal with an active warden Alice must embed her message in a robust manner. That is, Bob

Manuscript received June 7, 2001; revised September 26, 2002. This work was supported in part by TUBITAK BDP, Boğaziçi Research Fund Project 01A201, and NSF INT 9996097. N. Memon was supported by AFOSR Award Number F49620-01-1-0243. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christine Guillemot.

İ. Avcıbaş is with the Department of Electronics Engineering, Uludağ University, Bursa 16059, Turkey (e-mail: avcibas@uludag.edu.tr).

N. Memon is with the Department of Computer and Information Science, Polytechnic University, Brooklyn, NY 11201 USA (e-mail: memon@poly.edu).

B. Sankur is with the Department of Electrical and Electronics Engineering, Boğaziçi University, Istanbul, Turkey (e-mail: sankur@boun.edu.tr).

Digital Object Identifier 10.1109/TIP.2002.807363

should be able to accurately recover the secret message m despite operations like LSB randomizing, compression, filtering, and rotation by small degrees, etc. performed by the active warden Wendy. Indeed, the problem of embedding messages in a robust manner has been the subject of intense research in the image processing community, albeit for applications other than steganography, under the name of *robust digital watermarking* [3].

A robust digital watermark is an imperceptible signal added to digital content that can be later detected or extracted in order to make some assertion about the content. For example, the presence of her watermark can be used by Alice to assert ownership of the content. Recent years have seen an increasing interest in digital watermarking with many different applications, ranging from copyright protection and digital rights management, to secret communication. Essentially robust digital watermarks provide a means of image-based steganography in the presence of an active warden since modifications made by the warden will not affect the embedded watermark as long as the visual appearance of the image is not significantly degraded. However, despite this obvious and commonly observed connection to steganography, there has been very little effort aimed at analyzing or evaluating the effectiveness of common robust watermarking techniques for steganographic applications. Instead, most work has focused on analyzing or evaluating the watermarking algorithms for their robustness against various kinds of attacks that try to remove or destroy them. However, if robust digital watermarks are to be used in active warden steganography applications, detection of their presence by an unauthorized agent defeats their very purpose. Even in applications that do not require hidden communication, but only robustness, we note that it would be desirable to first detect the possible presence of a watermark before trying to remove or manipulate it. This means that a given signal would have to be first analyzed for the presence of a watermark.

In this paper, we develop steganalysis techniques both for conventional LSB-like embedding used in the context of a passive warden model and for watermarking which can be used to embed secret messages in the context of an active warden. In order to distinguish between these two models, we will be using the terms watermark and message when the embedded signal is in the context of an active warden and a passive warden, respectively. Furthermore, we simply use the terms marking or embedding when the context of discussion is general to include both active and passive warden steganography.

The techniques we present are novel and to the best of our knowledge, the first attempt at designing general purpose tools for steganalysis. General detection techniques as applied to steganography have not been devised and methods beyond visual inspection and specific statistical tests for individual techniques like LSB embedding [4]–[7] are not present in the literature. Since too many images have to be inspected visually to sense hidden messages, the development of a technique to automate the detection process will be very valuable to the steganalyst. Our approach is based on the fact that hiding information in digital media requires alterations of the signal properties that introduce some form of degradation, no matter how small. These degradations can act as signatures that could be used to reveal the existence of a hidden message. For example, in the context of digital watermarking, the general underlying idea is to create a watermarked signal that is *perceptually identical but statistically different* from the host signal. A decoder uses this statistical difference in order to detect the watermark. However, the very same statistical difference that is created could potentially be exploited to determine if a given image is watermarked or not. In this paper, we show that addition of a watermark or message leaves unique artifacts, which can be detected using Image Quality Measures (IQM).

The rest of this paper is organized as follows. In Section II, we discuss the selection of the image quality measures to be used in the steganalysis and the rationale for utilizing multiple quality measures. We then show that the image quality metric based distance between an *unmarked image* and its filtered version is different as compared to the distance between a *marked image* and its filtered version. Section III describes the regression analysis that we use to build a composite measure of quality to indicate the presence or absence of a mark. Statistical tests and experiments are given in Section IV and, finally, conclusions are drawn in Section V. The selected IQMs are described in the Appendix.

II. Choice of Image Quality Measures

The main goal of this paper is to develop a discriminator for cover images and stego images, using an appropriate set of IQMs. Image quality measurement continues to be the subject of intensive research and experimentation [8]–[11]. Objective image quality measures are based on image features, a functional of which, should correlate well with subjective judgment, that is, the degree of (dis)satisfaction of an observer [12]. Objective quality measures have been utilized in coding artifact evaluation, performance prediction of vision algorithms, quality loss due to sensor inadequacy etc. [13]. In this paper, however, we want to exploit image quality measures, not as predictors of subjective image quality or algorithmic performance, but specifically as a steganalysis tool, that is, as features in detecting watermarks or hidden messages.

A good IQM should be accurate, consistent and monotonic in predicting quality. In the context of steganalysis, *prediction accuracy* can be interpreted as the ability of the measure to detect the presence of hidden message with minimum error on average. Similarly, *prediction monotonicity* signifies that IQM scores should ideally be monotonic in their relationship to the embedded message size or watermark strength. Finally, *prediction consistency* relates to the quality measure's ability to provide consistently accurate predictions for a large set of watermarking or steganography techniques and image types. This implies that the spread of quality scores due to factors of image variety, active warden or passive warden steganography methods should not eclipse the score differences arising from message embedding artifacts. In order to understand how these metrics measure up to the above desiderata we resorted to analysis of variance (ANOVA) techniques. Specifically, ANOVA was used to show whether a metric's response was consistent with a change in the image or whether it was a random effect. The ranking of the goodness of the metrics was done according

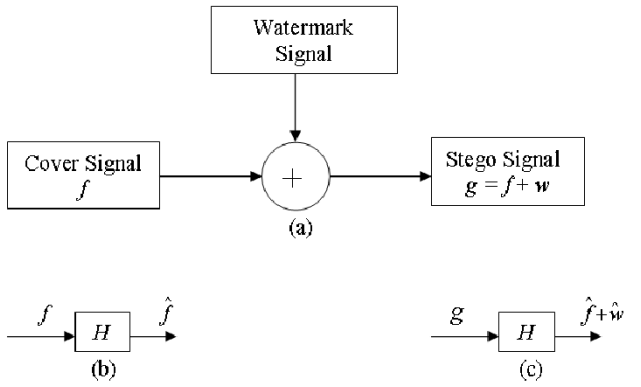


Fig. 1. Schematic descriptions of (a) watermarking or stegoing, (b) filtering an unmarked image, and (c) filtering a marked image.

to the F-scores in the ANOVA tests to identify the ones that responded most consistently and strongly. A similar study had been done in [11] to delineate good metrics to measure image quality under compression and sensor artifacts. In the final analysis we seek IQMs that are sensitive specifically to steganography effects, that is, those measures for which the variability in score data can be explained better because of some treatment rather then as random variations due to the image set.

The steganalysis detector we develop is based on regression analysis of a number of *relevant* IQMs. The idea behind detection of watermark or hidden message presence is to obtain a consistent distance metric for images containing a watermark or hidden message *vis-à-vis* those without, *with respect to a common reference*. The reference processing should possibly recover the original un-watermarked image, and to this purpose, we have used low-pass filtering based on a Gaussian kernel. In this respect other approaches such as denoising and Wiener filtering are also possible [14]. In fact Wiener filtering approach gave better results, for example, in the case of the Digimarc algorithm while denoising proved more effective in the case of Jsteg. However the Gaussian filtering approach was preferred because it gave uniformly good results across all steganographic techniques.

To clarify the rationale of our detector, let us recall that steganographic message embedding techniques, whether by spread-spectrum or quantization modulation or LSB insertion, can be represented as a signal addition to the cover image, as shown in Fig. 1. Let f be the cover image, $g = f + w$ be the stego-image, and w the inserted watermark. Let H be the ML (Maximum Likelihood) operator for the estimate of the watermark sequence. In the absence of any watermark or stego-signal $Hg = \hat{f}$ corresponds to the high-frequency content $\hat{f}$ of the image, while for a marked signal it yields $Hg = \hat{f} + \hat{w}$ where $\hat{w}$ denotes the ML estimate of the mark. The image quality metrics, in fact, are simply trained to differentiate between these two signals $\hat{f}$ and $\hat{f} + \hat{w}$. Fig. 2 gives an instance of the watermarked versus nonwatermarked class separability based on a scatter diagram of the three image quality metrics used. The training procedure for the steganalyzer is shown in Fig. 3(a).

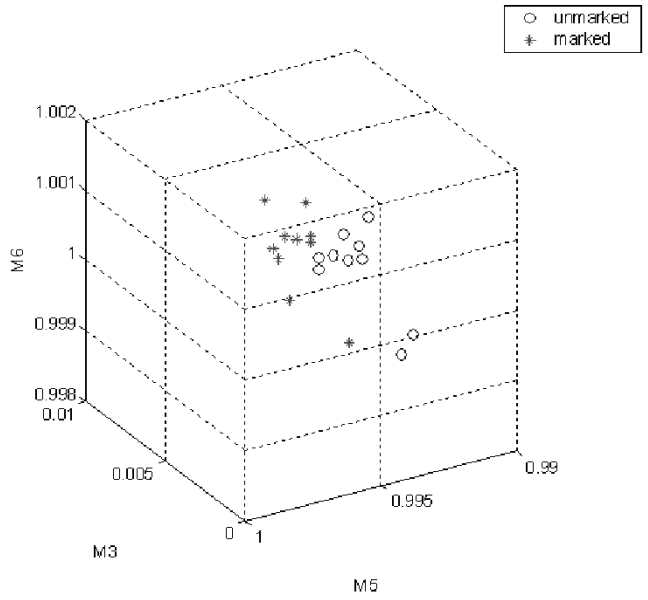


Fig. 2. Scatter plots of the three Image Quality Measures (M3: Czekakowski measure, M5: Image fidelity, and M6: Normalized cross-correlation).

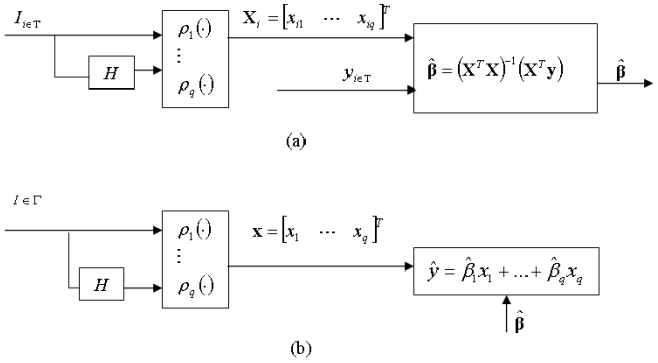


Fig. 3. Schematic description of (a) training and (b) testing.

The filter was chosen as a Gaussian smoothing filter $H(m, n) = Kg(m, n)$ where $g(m, n) = (2\pi\sigma^2)^{-1}\exp\{-(m^2 + n^2)/2\sigma^2\}$ is the 2-D Gaussian kernel and $K = (\sum_m \sum_n |g(m, n)|^2)^{-1/2}$ is the normalizing constant. The aperture of the Gaussian filter was set experimentally to $\sigma = 0.5$ with a mask size 3×3. The reason why Gaussian blurring works fine as a common reference is that it gives us the local mean which is also the maximum likelihood (ML) estimate of the image under Gaussian assumption [14]. Under a Laplacian distribution assumption the median would have been the ML estimate. Therefore the blurred image minus the original image yields the maximum likelihood estimate of the additive watermark. For the two ML estimators that we have tested, H is equivalent to the subtraction from the received stego-image of its local mean or median. Finally in the comparison between the mean and median filters as the ML estimates of the image we have found out that the former performs slightly better in the detection tests.

As for the selection of quality measures we used the results of a previous study [11] where several (26 in total) measures

were investigated to predict compression, blur and noise artifacts. From these measures we gleaned out the ones that served well the purpose of our steganalysis. The rationale of using several quality measures is that different measures respond with differing sensitivities to artifacts and distortions. For example, measures like mean-square-error respond more to additive noise, whereas others such as spectral phase or mean square HVS-weighted (Human Visual System) error are more sensitive to pure blur; while the gradient measure reacts to distortions concentrated around edges and textures. Similarly embedding techniques affect different aspects of images. In fact some watermarking algorithms inject "noise" in block DCT coefficients, others in a narrow-band of global DCT or Fourier coefficients, still others operate in selected localities in the spatial domain. Since we want our steganalyzer to be able to work with a variety of watermarking and steganographic algorithms, a multitude of quality features are needed so that the steganalyzer has the chance to probe several features in an image that are significantly impacted by the embedding process.

In order to identify specific quality measures that are useful in steganalysis, we used ANOVA [15] tests, with the expectation that it would distinguish measures that are consistent and accurate *vis-à-vis* the effects of watermarking and of steganography. More specifically several quality measures were statistically tested to determine if their fluctuations resulted from image variety or whether they were due to treatment effects of message embedding. ANOVA was used to show whether the variation in the data could be accounted for by the hypothesized factor, for example, the strength factor of watermarking or steganography. The hypotheses for the comparison of independent groups are

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{means of all the groups are equal,}$$

$$H_A: \quad \mu_i \neq \mu_j \quad \text{means of the two or more groups are not equal.}$$

It should be noted that the test statistic is an F test with $k-1$ and $N-k$ degrees of freedom, where N is the total number of watermarked or stegoed images. A low p-value (high F value) for this test indicates evidence for rejecting the null hypothesis in favor of the alternative. In other words, there is evidence that at least one pair of means are not equal. We opted to carry out the multiple comparison tests at a significance level of 0.05. Thus any test resulting in a p-value under 0.05 was considered to be significant, and therefore, one would reject the null hypothesis in favor of the alternative hypothesis. This is to assert that the difference in the quality metric arises from the "strength" parameter of the watermarking or steganography artifacts, and not from variations in the image content.

We performed three different ANOVA tests: The first was for active warden steganography, the second for passive warden steganography, and the last one for both active and passive warden steganography.

For active warden image tests, the first group consisted of the IQM scores computed from plain images and their filtered versions. The remaining three groups consisted of the IQM scores computed from watermarked images by Digimarc [16], PGS [17] and Cox [18] techniques, respectively, and their filtered versions. The data given to the ANOVA algorithm consisted of four vectors, each of dimension N, where $N = 12$ is the number of images used in the test from the training set. More specifically, consider a typical quality measure, say $M(\mu_i)$, where the parametric dependence upon the watermarking algorithm is shown with μ_i, $i = 0 \cdots 3$, for plain images, Digimarc, PGS and Cox techniques, respectively. The N-dimensional vector M reads as: $M(\mu_i) = [M(1|\mu_i) \cdots M(N|\mu_i)]^T$.

For passive warden image tests, the first group consisted of the IQM scores computed from plain (nonmarked) images, while the remaining three groups consisted of the IQM scores computed from images marked by Steganos [19], Stools [20] and Jsteg [21], respectively, and their filtered versions.

For the joint active warden and passive warden steganography analysis, the first group consisted of the IQM scores computed from plain images. The remaining six groups consisted of the IQM scores computed from watermarked images by Digimarc, PGS and Cox technique, marked images by Steganos, Stools, and Jsteg, respectively, and their respective filtered versions.

TABLE I
ONE-WAY ANOVA TESTS FOR WATERMARKING, STEGANOGRAPHY, AND POOLED WATERMARKING AND STEGANOGRAPHY

	Watermark		Stego		Watermark & Stego	
Image Quality Measures	F	p	F	p	F	p
Minkowsky Metric $\gamma = 2$	**6.06**	0.01	0.56	0.58	5.28	0.00
Minkowsky Metric $\gamma = 1$	**3.28**	0.05	0.57	0.58	**3.07**	0.03
Maximum Difference	0.13	0.93	0.31	0.74	0.25	0.93
Sorted Maximum Difference	0.14	0.93	0.07	0.92	0.13	0.98
Czekanowski	**4.63**	0.02	1.08	0.37	**4.66**	0.01
Structural Content	0.62	0.61	0.15	0.86	0.58	0.71
Cross Correlation	2.08	0.14	0.21	0.81	0.74	0.60
Image Fidelity	2.67	0.08	0.40	0.68	1.14	0.37
Angle Mean	1.95	0.17	**4.20**	0.04	**3.40**	0.02
Angle Standard Deviation	0.45	0.72	3.27	0.08	2.36	0.08
Spectral Magnitude	**5.50**	0.03	0.02	0.98	**4.35**	0.01
Spectral Phase	**5.49**	0.03	0.02	0.98	4.34	0.01
Weighted Spectral Distance	1.12	0.37	0.06	0.94	0.66	0.65
Median Block Spectral Magnitude	0.79	0.51	0.001	0.99	0.44	0.81
Median Block Spectral Phase	0.47	0.72	**3.95**	0.05	**4.24**	0.02
Median Block Weighted Spectral Distance	0.45	0.72	**3.96**	0.05	**4.22**	0.02
Normalized Absolute Error (HVS)	0.16	0.92	1.16	0.35	0.74	0.61
Normalized Mean Square ERROR (HVS)	**3.30**	0.05	**4.93**	0.02	**2.69**	0.05
HVS Based L2	0.19	0.90	0.46	0.64	0.47	0.79

In Table I we give ANOVA results with respect to active warden, passive warden and combined techniques. The measures that have higher discriminative power—measures that catch the statistical evidence of steganography—are shown in bold. These measures, in fact, sense better the statistical difference between the populations of marked and nonmarked images so that they can be used to separate the two classes. The implications of the result are twofold. One is that, using these features a steganalysis tool can be designed to detect marked images, as we show in Section III, using multivariate regression analysis. The other is that, current steganographic algorithms should exercise more care on these statistically significant image features to eschew detection. It is interesting to note that the significance ordering of the IQMs for active warden and passive warden steganographic algorithms are different. For instance while the Minkowsky measures were not statistically significant for passive warden steganographic algorithms, they were for the active warden algorithms. Minimizing the Mean Square Error (MSE) or the Kullback–Leibler distance between

the original (cover) image and the stego image is not necessarily enough to achieve covert communication as the evidence can be caught by another measure such as spectral measures. The selected subset of image quality measures in the design of steganalyzer with respect to their statistical significance were as follows.

Active Warden Steganography: Mean Absolute Error M_1, Mean Square Error M_2, Czekznowski Correlation Measure M_3, Image Fidelity M_5, Cross Correlation M_6, Spectral Magnitude Distance M_7, Normalized Mean Square HVS Error M_{10}. We denote this feature set as $\Psi = \{M_1, M_2, M_3, M_5, M_6, M_7, M_{10}\}$ for future reference in the experiments in Section IV.

Passive Warden Steganography: Angle Mean M_4, Median Block Spectral Phase Distance M_8, Median Block Weighted Spectral Distance M_9, Normalized Mean Square HVS Error M_{10}. We denote this feature set as $\Omega = \{M_4, M_8, M_9, M_{10}\}$.

Pooled Active Warden and Passive Warden Steganography: Mean Absolute Error M_1, Mean Square Error M_2, Czekanowski Correlation Measure M_3, Angle Mean M_4, Spectral Magnitude Distance M_7, Median Block Spectral Phase Distance M_8, Median Block Weighted Spectral Distance M_9, Normalized Mean Square HVS Error M_{10}. We denote this feature set as $\Xi = \{M_1, M_2, M_3, M_7, M_8, M_9, M_{10}\}$.

III. Regression Analysis of the Quality Measures

The steganalysis we propose is based on the observation in Section II that an embedded and filtered image differs statistically from a nonembedded but simply filtered image. This statistical difference can be put in light by comparing the embedded image and its original version against a common reference treatment that is their filtered versions. It has been observed that filtering an image with no watermarked message causes changes in the IQMs differently than the changes brought about on embedded images. This differential behavior is in part because steganographic embedding is not in general a global operation, but is local in nature. The message signal is either injected locally, e.g., on a block basis, or the signal is subjected to a perceptual mask. In any case, we consistently obtained statistically different quality scores from embedded-and-filtered images and from filtered-but-not-embedded sources. For the hypothesis testing we used the quality scores, which are separately calculated for differences obtained from a nonembedded image and its embedded varieties.

In the design phase of the steganalyzer, we regressed the normalized IQM scores to, respectively, -1 and 1, depending upon whether an image did not or did contain a message. Similarly, IQM scores were calculated between the original images and their filtered versions. In the regression model [15], we expressed each decision label y in a sample of n observations as a linear function of the IQM scores, denoted as x's, plus a random error, ε

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_q x_{1q} + \varepsilon_1 \\ y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_q x_{2q} + \varepsilon_2 \\ &\vdots \\ y_N &= \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_q x_{Nq} + \varepsilon_N. \end{aligned} \tag{1}$$

In this expression, x_{ij} denotes the IQM score, where the first index indicates the ith image $i = 1, \ldots, N$, and the second one the quality measure, $j = 1, \ldots, q$, q being the total number of quality measures considered. The βs denote the regression coefficients. The complete statement of the standard linear model is

$$y = X_{Nxq}\beta + \varepsilon \tag{2}$$

where the Nxq data matrix has rank q, and ε is a zero-mean Gaussian noise. The corresponding optimal MMSE linear predictor $\boldsymbol{\beta}$ can be obtained by

$$\hat{\beta} = (X^T X)^{-1}(X^T y). \tag{3}$$

Once the prediction coefficients are obtained in the training phase, these coefficients can be used in the testing phase. Given an image in the test phase, first it is filtered and the q IQM scores are obtained using the image and its filtered version. Then using the prediction coefficients, these scores are regressed to the output value. If the output exceeds the threshold 0 then the decision is that the image is *embedded*, otherwise the decision is for *not embedded*. That is

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_q x_q \tag{4}$$

for $\hat{y} \geq 0$ the image contains watermark, and for $\hat{y} < 0$ it does not. The schematic diagram of the steganalyzer in the test phase is given in Fig. 3(b).

IV. Simulation Results

The active warden techniques we used were the following: Photoshop plug-in Digimarc [16], Cox's technique [18], and the technique from Swiss Federal Institute of Technology, PGS [17]. One obvious reason for selecting the above techniques was their free availability on the Internet and that they were all popularly known algorithms. A more relevant reason was that these techniques permitted adjusting the watermark insertion strength, which was instrumental to probe the sensitivity of IQMs. On the other side the three passive warden steganographic tools selected were Steganos [19], S-Tools [20] and Jsteg [21]. These tools were among the most cited ones for their satisfactory results with respect to steganographic applications. We used an image database from [22] for the simulations. The database contained an adequate variety of images including computer generated images, images with bright or with reduced and dark colors, images with textures and fine details, and some well-known images like Lena, peppers etc. We performed eight experiments organized in three sets.

The three experiments 1)–3) in the first set involved active warden techniques only, namely: 1) First, the individual steganalysis of each watermarking algorithm, Digimarc, PGS and Cox at admissible watermark strengths; 2) Second, the steganalysis of pooled watermarking algorithms at admissible watermark strengths; 3) Third, the cross-validation experiment where the steganalyzer was trained on images watermarked by Digimarc, and tested on images watermarked by PGS and Cox *et al.*

The next three experiments 4)–6) in the second set involved passive warden steganography only: 4) The steganalysis of individual steganography algorithms, Steganos, Stools and Jsteg for different embedded message sizes; 5) The steganalysis of pooled steganography algorithms for different message sizes;

TABLE II
TRAINING AND TEST SAMPLES FOR DIGIMARC AND PGS FOR EXPERIMENT 1

	Level 1	Level 2	Level 3	Level 4
Training samples	1,2,3	4,5,6	7,8,9	10,11,12
Test samples	13,14,15	16,17	18,19,20	21,22

TABLE III
TRAINING AND TEST SAMPLES FOR COX FOR EXPERIMENT 1

	1000 coefficients
Training samples	1…12
Test samples	13…22

TABLE IV
TRAINING AND TEST SAMPLES FOR POOLED WATERMARKING ALGORITHMS FOR EXPERIMENT 2 (L1: LEVEL 1 ETC.)

	Digimarc				PGS				Cox
WM Levels	L1	L2	L3	L4	L1	L2	L3	L4	
Training samples	1	2	3	4	5	6	7	8	9,10,11,12
Test samples	13	14	15	16	17	18	19	20	21,22

TABLE V
TRAINING AND TEST SAMPLES FOR EXPERIMENT 3: TRAIN ON DIGIMARC, TEST ON PGS AND COX

Training	Digimarc			
WM Levels	L1	L2	L3	L4
Training samples	1…3	4…6	7…9	10…12
Testing	PGS		COX	
WM Levels	L1	L2	L3	
Test samples	13…15	16…18	19,20	21,22

6) In the sixth experiment the steganalyzer was trained on images embedded with Steganos and Stools, and tested on images embedded with Jsteg for cross-validation purposes.

In the third set the final two experiments 7) and 8) involved both active warden and passive warden steganography algorithms. The seventh experiment was steganalysis of the pooled three passive warden and three active warden steganographic algorithms for admissible levels of watermark strength and for different message lengths. In the last and eighth experiment the steganalyzer was trained on images embedded with Steganos, Stools, or watermarked by Digimarc and tested on images embedded with Jsteg or watermarked by Cox *et al.* The aim of the last two experiments, in the same spirit as in experiments 3) and 6), was to see the generalizing ability of the steganalyzer in case an image was to be marked with a method unknown to it in the learning phase. In experiments 1)–3) the feature set was Ψ which was defined in Section II, for the experiments 4)–6) the feature set was Ω, while the feature set was Ξ for the remaining experiments 7) and 8).

The organizations of the training and testing samples for the experiments are given in Tables II–XII. The images in the training and test sets are denoted by numbers. More specifically the training set is $T = \{1, \ldots, 12\}$ and the test set is $\Gamma = \{13, \ldots, 22\}$. There were four levels of watermark strength for Digimarc and PGS (denoted by L1 to L4 in the Tables). We used the original settings of Cox's technique; modified the 1000 most significant coefficients in spectral domain. The embedded message sizes were 1/10 and 1/40 of the cover image size for Steganos and Stools, while the message sizes were 1/100 of the cover image size for Jsteg.

TABLE VI
TRAINING AND TEST SAMPLES FOR STOOLS FOR EXPERIMENT 4

Message size	1/40 of image size	1/10 of image size
Training samples	1…6	7…12
Test samples	13…17	18…22

TABLE VII
TRAINING AND TEST SAMPLES FOR JSTEG FOR EXPERIMENT 4

Message size	1/100 of image size
Training samples	1…12
Test samples	13…22

TABLE VIII
TRAINING AND TEST SAMPLES FOR STEGANOS FOR EXPERIMENT 4. (NOTE: IN CERTAIN IMAGES THE STEGANOS DID NOT LET THE MESSAGES TO BE EMBEDDED NO MATTER WHAT THEIR SIZE)

Message size	1/40 of image size	1/10 of image size
Training samples	2,4,8	10,11,13
Test samples	15,17	19,20,21

TABLE IX
TRAINING AND TEST SAMPLES FOR POOLED STEGANOGRAPHY ALGORITHMS FOR EXPERIMENT 5

	Steganos		Stools		Jsteg
Message size	1/40	1/10	1/40	1/10	1/100
Training samples	2,4	8,10	1,3	5,6	7,9,11,12
Test samples	13,15	17,19	14,16	18,20	21,22

TABLE X
TRAINING AND TEST SAMPLES FOR EXPERIMENT 6: TRAIN ON STEGANOS AND STOOLS, TEST ON JSTEG

Training	Steganos		Stools	
Msg. Size	1/40	1/10	1/40	1/10
Training samples	2,4,8	10,11	1,3,5,6	7,9,12
Testing	Jsteg			
Msg. Size	1/100			
Test samples	13…22			

TABLE XI
TRAINING AND TEST SAMPLES FOR POOLED WATERMARKING AND STEGANOGRAPHY ALGORITHMS FOR EXPERIMENT 7

	Digimarc		PGS		Cox	Steganos		Stools		Jsteg
Level or msg size	L2	L3	L2	L3	1000 cof	1/40	1/10	1/40	1/10	1/100
Training samples	7	8	9	10	11,12	2	4	1	3	5,6
Test samples	18	19	20	21	22	13	15	14	16	17

TABLE XII
TRAINING AND TEST SAMPLES FOR EXPERIMENT 8: TRAIN ON STEGANOS, STOOLS AND DIGIMARC, TEST ON JSTEG AND COX

Training	Digimarc		PGS		Cox	Steganos		Stools		Jsteg
Level or msg size	L2	L3	L2	L3	1000 cof	1/40	1/10	1/40	1/10	1/100
Training samples	7	9	11	12		2,4	8,10	1,3	5,6	
Testing	Digimarc		PGS		Cox	Steganos		Stools		Jsteg
Level or msg size	L2	L3	L2	L3	1000 cof	1/40	1/10	1/40	1/10	1/100
Test samples					13…17					18..22

The performance of the steganalyzer is given in Table XIII. Simulation results indicate that the selected IQMs form a multidimensional feature space whose points cluster well enough to do a classification of marked and nonmarked images. The classifier is still able to do a classification when the tested images come from an embedding technique unknown to it, indicating that it has a generalizing capability of capturing the general intrinsic characteristics of steganographic techniques.

TABLE XIII
PERFORMANCE OF THE STEGANALYZER FOR ALL THE EXPERIMENTS

Experiment	False Alarm	Miss	Correct Detection	Per. %
1. a. Digimarc	2/10	2/10	16/20	80
1. b. PGS	2/10	1/10	17/20	85
1. c. Cox	4/10	2/10	14/20	70
2. Pooled Watermarking	3/10	3/10	14/20	70
3. Train on Digimarc, Test on PGS and Cox	5/10	2/10	13/20	65
4. a. Steganos	2/5	1/5	7/10	70
4. b. Stools	4/10	1/10	15/20	75
4. c. Jsteg	3/10	3/10	14/20	70
5. Pooled Steganography	5/10	0/10	15/20	75
6. Train on Steganos and Stools, Test on Jsteg	3/10	3/10	14/20	70
7. Pooled Watermarking and Steganography	5/10	1/10	14/20	70
8. Train on Digimarc, PGS, Steganos, Stools Test on Cox and Jsteg	4/10	3/10	13/20	65

It may be argued that the statistical classification scores leave something to be desired. We would like to point out, however, that our goal was to design a general steganalysis tool that would perform adequately across several techniques. Certainly the performance of the steganalysis algorithm can be improved by constraining the domain and the set of algorithms. In fact recent years have seen many steganalysis techniques proposed in the literature such as [4], [5], [7]. The proposed algorithm is more general, however, in that it does not assume only spatial or only spectral domain embedding.

V. CONCLUSIONS

In this paper, we have addressed the problem of steganalysis of images, and we have developed a technique for discriminating between cover-images and stego-images. Our approach is based on the hypothesis that message-embedding schemes leave statistical evidence or structure in images that can be exploited for detection. In fact we have shown that the distance in the feature space between an unmarked and a reference image is different than the distance between a marked image and its reference version. We used image quality metrics as the feature set. To identify good features (quality measures), which provide the best discriminative power, we used ANOVA technique. A different point of view of the IQM-based steganalysis would be that these very image features should be taken into account in the design of watermarking or steganographic techniques if eschewing detection is desired. After selecting an appropriate feature set, we used multivariate regression techniques to get an optimal classifier. Simulation results with well known and commercially available watermarking and steganographic techniques indicate that the selected IQMs form a multidimensional feature space whose points cluster well enough to do a classification of marked and nonmarked images. The classifier is still able to do a classification when the tested images come from an embedding technique unknown to it, indicating that it has a generalizing capability of capturing the general intrinsic characteristics of watermarking and steganographic techniques. Future work will expand, on the one hand, the scope of the algorithm (the type of watermark algorithms, the media such as audio) and, on the other hand, to improve its detection performance, e.g., via decision fusion.

APPENDIX

We give brief descriptions of the selected image quality measures in this Appendix. In Table I, 19 IMQs are quoted, but here we describe the 10 selected measures that qualify in the ANOVA tests (indicated in bold characters in the Table). We denote multispectral components of an image at the pixel position i, j, and in band k as $C_k(i, j)$, where $k = 1, \ldots, 3$ for color images. The boldface symbols, $\mathbf{C}(i, j)$, $\hat{\mathbf{C}}(i, j)$ indicate the multispectral pixel vectors at position (i, j). The multiband image matrix is denoted by $\mathbf{C}$and $\hat{\mathbf{C}}$, where the hat superscripted quantity is the distorted (e.g., watermarked) version of the image. We will use M_i, $i = 1 \cdots 10$ to describe the ten IMQ features used in the detector.

A. Minkowsky Measures

The L_γ norm of the dissimilarity of two images can be calculated by taking the Minkowsky average of the pixel differences spatially and then chromatically (that is over the bands)

$$M_\gamma = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{N^2}\sum_{i,j=1}^{N}\left|C_k(i,j)-\hat{C}_k(i,j)\right|^\gamma\right\}^{1/\gamma}. \quad (5)$$

$\gamma = 1$ corresponds to mean absolute error (M_1), and $\gamma = 2$ to mean square error M_2, respectively.

B. Correlation Measures

A measure to compare vectors with strictly nonnegative components, as in the case of images, is the Czekanowski distance [23]

$$M_3 = \frac{1}{N^2}\sum_{i,j=0}^{N-1}\left(1-\frac{2\sum_{k=1}^{K}\min\left(C_k(i,j),\hat{C}_k(i,j)\right)}{\sum_{k=1}^{K}\left(C_k(i,j)+\hat{C}_k(i,j)\right)}\right). \quad (6)$$

A variant of correlation-based measures is the statistics of the angles between the pixel vectors of the two images. Similar colors will result in vectors pointing in the same direction, while significantly different colors will point in different directions in the color space. Since we deal with positive vectors $\mathbf{C}$, $\hat{\mathbf{C}}$, we are constrained to the first quadrant of the Cartesian space so that the maximum difference attained will be $\pi/2$. The angular correlation between two vectors is defined as follows [24]:

$$M_4 = 1-\frac{1}{N^2}\sum_{i,j=1}^{N}\frac{2}{\pi}\cos^{-1}\frac{\left\langle\mathbf{C}(i,j),\hat{\mathbf{C}}(i,j)\right\rangle}{\|\mathbf{C}(i,j)\|\left\|\hat{\mathbf{C}}(i,j)\right\|}. \quad (7)$$

The closeness between two digital images can also be quantified in terms of correlation function. The Image Fidelity and Normalized Cross-Correlation measures are defined, respectively, as follows:

$$M_5 = 1-\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i,j=0}^{N-1}\left[C_k(i,j)-\hat{C}_k(i,j)\right]^2}{\sum_{i,j=0}^{N-1}C_k(i,j)^2}\right), \quad (8)$$

$$M_6 = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i,j=0}^{N-1}C_k(i,j)\hat{C}_k(i,j)}{\sum_{i,j=0}^{N-1}C_k(i,j)^2}. \quad (9)$$

C. Spectral Measures

Let the Discrete Fourier Transforms (DFT) of the kth band of the original and embedded image be denoted by $\Gamma_k(u, v)$ and $\hat{\Gamma}_k(u, v)$, respectively. The spectra are defined as

$$\Gamma_k(u, v) = \sum_{m,n=0}^{N-1} C_k(m, n) \exp\Big[-2\pi i m \frac{u}{N}\Big] \exp\Big[-2\pi i n \frac{v}{N}\Big], \quad k = 1 \cdots K. \tag{10}$$

The phase and magnitude spectra are defined, respectively as $\varphi(u, v) = \arctan(\Gamma(u, v))$ and $M(u, v) = |\Gamma(u, v)|$. The spectral magnitude distortion measure is given by

$$M_7 = \frac{1}{KN^2} \sum_{k=1}^{3} \sum_{u,v=0}^{N-1} \Big| |\Gamma_k(u, v)| - \Big|\hat{\Gamma}_k(u, v)\Big| \Big|^2. \tag{11}$$

Due to the localized nature of distortion and/or the nonstationary image field, Minkowsky averaging of block spectral distortions may be more advantageous. Thus an image can be divided into L blocks of size $b \times b$, say 32×32, and block wise spectral distortions can be computed. Let the DFT of the lth block of the kth band image $C_k^l(m, n)$ be $\Gamma_k^l(u, v)$

$$\Gamma_k^l(u, v) = \sum_{m,n=0}^{b-1} C_k^l(m, n) \exp\Big[-2\pi i m \frac{u}{b}\Big] \exp\Big[-2\pi i n \frac{v}{b}\Big] \tag{12}$$

where $u, v = -(b/2) \cdots (b/2)$ and $l = 1, \ldots, L$, or in the magnitude-phase form

$$\Gamma_k^l(u, v) = \Big|\Gamma_k^l(u, v)\Big| e^{j\phi_k^l(u, v)}. \tag{13}$$

Then the following measures can be defined in the transform domain over the lth block

$$J_M^l = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{u,v=0}^{b-1} \Big(\big|\Gamma_k^l(u, v)\big| - \Big|\hat{\Gamma}_k^l(u, v)\Big| \Big)^{\gamma} \right)^{1/\gamma}, \tag{14}$$

$$J_\varphi^l = \frac{1}{K} \sum_{k=1}^{K} \left(\sum_{u,v=0}^{b-1} \Big(\big|\phi_k^l(u, v)\big| - \Big|\hat{\phi}_k^l(u, v)\Big| \Big)^{\gamma} \right)^{1/\gamma} \tag{15}$$

$$J^l = \lambda J_M^l + (1 - \lambda) J_\varphi^l \tag{16}$$

with λ the relative weighting factor of the magnitude and phase spectra. Among possible rank order operations on the block spectral differences the median has proven useful. The norm parameter set at $\gamma = 2$ and block size of 32×32 yielded higher F scores. Weighting parameter λ is chosen so as to render the contributions of the magnitude and phase terms commensurate. Median of block spectral phase and median of weighted block spectral distortion measures are defined, respectively, as

$$M_8 = \underset{l=1\cdots L}{median}\ J_\phi^l, \tag{17}$$

$$M_9 = \underset{l=1\cdots L}{median}\ J^l. \tag{18}$$

D. HVS Based Measure

The incorporation of human visual system (HVS) model into objective measures [25], [26] has led to a better correlation with the subjective ratings in multimedia. It is conjectured therefore that in steganalysis tasks they may have as well some relevance. We assume that the human visual system can be modeled as a band-pass filter with a transfer function in polar coordinates,

$$H(\rho) = \begin{cases} 0.05 e^{\rho^{0.554}} & \rho < 7 \\ e^{-9[|\log_{10}\rho - \log_{10} 9|]^{2.3}} & \rho \geq 7 \end{cases} \tag{19}$$

where $\rho = (u^2 + v^2)^{1/2}$. Once images are processed with such a spectral mask and inverse DCT transformed, the Normalized Mean Square HVS Error is defined as

$$M_{10} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{i,j=0}^{N-1} \Big[U\{C_k(i, j)\} - U\Big\{\hat{C}_k(i, j)\Big\} \Big]^2}{\sum_{i,j=0}^{N-1} \big[U\{C_k(i, j)\}\big]^2}. \tag{20}$$

REFERENCES

[1] G. J. Simmons, "'Prisoners' problem and the subliminal channel," in *CRYPTO83—Advances in Cryptology*, 1984, pp. 51–67.

[2] N. F. Johnson and S. Katzenbeisser, "A survey of steganographic techniques," in *Proc. Information Hiding*, Norwood, MA, 2000, pp. 43–78.

[3] J.-L. Dugelay and S. Roche, "A survey of current watermarking techniques," in *Information Hiding Techniques for Steganography and Digital Watermarking*, S. Katzenbeisser and F. A. P. Petitcolas, Eds. Norwood, MA: Artech House, 1999, ch. 6.

[4] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," in *Proc. 3rd Information Hiding Workshop*, Dresden, Germany, 1999, pp. 61–76.

[5] N. F. Johnson and S. Jajodia, "Steganalysis: The investigation of hidden information," in *Proc. IEEE Inform. Technol. Conf.*, Syracuse, NY, 1998.

[6] ——, "Steganalysis of images created using current steganography software," in *Proc. Workshop on Information Hiding*, ser. Lecture Notes in Computer Science. Portland, OR: Springer-Verlag, 1998, vol. 1525, pp. 273–289.

[7] J. Fridrich, R. Du, and M. Long, "Steganalysis of LSB encoding in color images," in *Proc. ICME 2000*, New York, 2000.

[8] A. M. Eskicioğlu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, pp. 2959–2965, Dec. 1995.

[9] A. M. Eskicioğlu, "Application of multidimensional quality measures to reconstructed medical images," *Opt. Eng.*, vol. 35, pp. 778–785, Mar. 1996.

[10] B. Lambrecht, Ed., "Special issue on image and video quality metrics," in *Signal Process.*, Oct. 1998, vol. 70.

[11] İ. Avcıbaş, B. Sankur, and K. Sayood, "Statistical analysis of image quality measures," *J. Electron. Imag.*, vol. 11, pp. 206–223, Apr. 2002.

[12] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–205.

[13] C. E. Halford, K. A. Krapels, R. G. Driggers, and E. E. Burroughs, "Developing operational performance metrics using image comparison metrics and the concept of degradation space," *Opt. Eng.*, vol. 38, pp. 836–844, May 1999.

[14] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The watermark copy attack," in *Proc. SPIE Conf. on Security and Watermarking of Multimedia Contents II*, San Jose, CA, 2000, pp. 371–380.

[15] A. C. Rencher, *Methods of Multivariate Analysis*. New York: John Wiley, 1995, ch. 6, 10.

[16] PictureMarc, Embed Watermark, v 1.00.45, Digimarc Corp., .

[17] M. Kutter and F. Jordan. JK-PGS (Pretty Good Signature). [Online]. Available: http://ltswww.epfl.ch/~kutter/watermarking/JK_PGS.html.

[18] J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vol. 6, pp. 1673–1686, Dec. 1997.

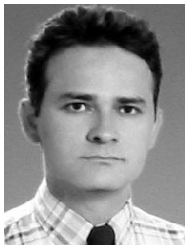
[19] Steganos II Security Suite.. [Online]. Available: http://www.steganos.com/english/steganos/download.htm.

[20] A. Brown. S-tools version 4.0. [Online]. Available: http://members.tripod.com/steganography/stego/s-tools4.html.

[21] J. Korejwa. Jsteg shell 2.0. [Online]. Available: http://www.tiac.net/users/korejwa/steg.htm.

[22] Images. [Online]. Available: http://www.cl.cam.ac.uk/~fapp2/watermarking/benchmark/image_database.html.

[23] Percentage similarity, (syn. Czekanowski coefficient). [Online]. Available: http://ag.arizona.edu/classes/rnr555/lecnotes/10.html.
[24] P. E. Trahanias, D. Karakos, and A. N. Venetsanopoulos, "Directional processing of color images: Theory and experimental results," *IEEE Trans. Image Processing*, vol. 5, pp. 868–880, June 1996.
[25] A. B. Watson, Ed., *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993.
[26] N. B. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, vol. 33, pp. 551–557, June 1985.



İsmail Avcıbaş (M'02) received the B.S. and M.S. degrees in electronics engineering from Uludağ University, Bursa, Turkey, in 1992 and 1994, respectively, and the Ph.D. degree in electrical and electronics engineering from Boğaziçi University, İstanbul, Turkey, in 2001.

He received a scholarship from The Scientific Council of Turkey TUBITAK, BDP Program, and did research on image compression and steganalysis in the Department of Computer and Information Science, Polytechnic University, Brooklyn, NY, in 1999–2000. He is currently a Lecturer with the Department of Electronics Engineering, Uludağ University, Bursa, Turkey. His current research interests are in signal processing, data compression, steganalysis of audio–visual data, and pattern recognition.

Nasir Memon (M'92) is an Associate Professor in the Computer Science Department at Polytechnic University, New York. His research interests include data compression, computer security, and multimedia communication and computing. He has published more than 100 articles in journals and conference proceedings and holds two patents in image compression. He has been the Principal Investigator on several funded research projects sponsored by NSF as well as industry. He was a Visiting Faculty at Hewlett-Packard Research Labs during the academic year 1997–1998. He is an associate editor for *ACM Multimedia Systems Journal*.

Dr. Memon is currently an associate editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is also a guest editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING special issue on multimedia security.

Bülent Sankur (M'76–SM'90) received the B.S. degree in electrical engineering from Robert College, Istanbul, Turkey, and the M.Sc. and Ph.D. degrees from Rensselaer Polytechnic Institute, Troy, NY.

He has been active at Boğaziçi University in the Department of Electric and Electronic Engineering in establishing curricula and guiding research in the areas of digital signal processing, image and video compression, and multimedia systems. He has held visiting positions at the University of Ottawa, Canada; Istanbul Technical University; Technical University of Delft, The Netherlands; and Ecole Nationale Supérieure des Telecommunications, France.