A Classifier Design For Detecting Image Manipulations

Ismail Avcibas, Sevinc Bayram, Nasir Memon, Mahalingam Ramkumar, Bulent Sankur

Department of Electronics Engineering, Uludag University, Bursa, Turkey.
Department of Computer and Information Science, Polytechnic University, Brooklyn, NY, USA.
Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Turkey.
Department of Computer Science, Mississippi State University, Jackson, MS , USA.

Abstract

In this paper we present a framework for digital image forensics. Based on the assumptions that some processing operations must be done on the image before it is doctored, and an expected measurable distortion after processing an image, we design classifiers that discriminates between original and processed images. We propose a novel way of measuring the distortion between two images, one being the original and the other processed. The measurements are used as features in classifier design. Using these classifiers we test whether a suspicious part of a given image has been processed with a particular method or not. Experimental results show that with a high accuracy we are able to tell if some part of an image has undergone a particular or a combination of processing methods.

1. Introduction

In today's digital age, the creation and manipulation of digital images is made simple by digital processing tools that are easily and widely available. As a consequence, we can no longer take the authenticity of images for granted especially when it comes to legal photographic evidence. *Image forensics*, in this context, is concerned with determining the source and potential authenticity of an image.

Although digital watermarks have been proposed as a tool to provide authenticity to images, it is a fact that the overwhelming majority of images that are captured today do not contain a digital watermark. And this situation is likely to continue for the foreseeable future. Hence in the absence of widespread adoption of digital watermarks, there is a strong need for developing techniques that can help us make statements about the origin, veracity and authenticity of digital images.

In this paper we focus on the problem of reliably discriminating between "doctored" images (images which are altered in order to deceive people) from untampered original ones. The basic idea behind our approach is that a doctored image (or the least parts of it) would have undergone some image processing operations like scaling, rotation, brightness adjustment etc. Hence we first design classifiers that can distinguish between images that have and have not been processed using these basic operations. Then equipped with these classifiers we apply them successively to a suspicious sub-image of a target image and classify the target as doctored if a sub-image classifies differently from the rest of the image.

The rest of this paper is organized as follows: In Section 2 we present a method to compute content independent distortion measure that are used as features in the classifier we design for image forensics. Statistical performance results are given in Section 3, with conclusions drawn in Section 4.

2. Content Independent Features

Our goal is to design a feature based classifier that can discriminate between doctored and original images. The features we use for the classifier should be such that they reflect the distortions an image suffers from manipulation. A classifier based on these statistical features would then differentiate between the two cases of original versus doctored images, even when casual observers cannot perceive them visually. In this section we present a technique for capturing image features that, under some assumptions, are independent of original image content and hence better represent image manipulations.

Now, a doctored image could have been subjected to many operations like scaling, rotation, brightness adjustment, blurring, enhancement etc. or some particular combination thereof. Often doctoring may also involve cutting and pasting of another sub-image, which is skillfully manipulated and rendered along the suture into the original to avoid any suspicion. Since image manipulations can be very subtle, the discriminating features one employs can easily be overwhelmed by variations in the image content.

Keeping the above points in mind it is important to obtain features that remain independent of the image content, so that they would only reflect the presence, if any, of image manipulations. This is due to the fact that in any feature based classification method, there is the risk that the variability in the image content itself may eclipse image alterations present from the detector. Thus, it is desired that whatever features are selected, the detector respond only to *the induced distortions* during doctoring, and not be confused by the statistics of the image content.

In a previous study, we had shown the potential of certain image quality metrics in predicting the presence of steganographic signals within an image [2, 1]. Similar to this approach, we employ mul1le image quality metrics as the underlying features of our classifier. The rationale for using mul1le quality metrics is to probe different quality aspects of the image, which could be impacted during doctoring manipulations. For example, some measures respond at pixel level, others at the block level, yet others to edge distortions or spectral phase distortion.

Now the main reason image dependence creeps into the classifier is due to the fact that the original image (ground-truth) obviously is not available during the testing stage. Therefore some "ground-truth" or reference signal must be created common to both the training and testing stages. In our previous work on image steganalysis [1], we used a denoised version of the given image as the ground-truth reference. However, creating a reference signal via its own denoised version is obviously a content-dependent scheme.

In the rest of this section, we present an approach to preclude content dependency, by employing a reference image in the feature extraction process. More specifically, let x denote a test image and $x+\varepsilon$ be its processed version, and similarly let y and $y+\eta$ indicate the reference image and is processed version. Furthermore, consider a generic distortion functional $M(a,b)$ between two signals a and b. A simple example of which being the well known mean-square distortion function, $M(a,b) \stackrel{\Delta}{=} E[(a-b)^2]$, with E being the expectation operator. The classifier we design will be based on the statistics of the difference of the distortions, as will be explained in the sequel.

We now make two assumptions for the operation of our classifier. First, we assume the processing operations involved in image doctoring lead to additive distortion, i.e., that is, the altered signals can be represented as $x+\varepsilon$ and $y+\eta$. Second, we assume the additive distortions of the test and reference images are not mutually orthogonal, that is, $E\{\varepsilon^* \eta\} \neq 0$.

We first show that self-referencing, as employed in [1] causes content-dependent distortion. Let f be the specific operation to obtain the reference image; for example in [1] we used a denoising operation. In other words, we had $y = f(x) = denoise(x)$. The outcomes of this operation are given by $x \xrightarrow{f} f(x)$, $x+\varepsilon \xrightarrow{f} f(x+\varepsilon)$, respectively, for original signal and its processed version. To illustrate the point, for the case of the mean-square distortion one obtains:

$$M(x+\varepsilon, f(x+\varepsilon)) - M(x, f(x)) =$$
$$E[f(x+\varepsilon)^2 + 2x\varepsilon + \varepsilon^2 -$$
$$2(x+\varepsilon) f(x+\varepsilon) - 2x f(x) - f(x)^2] \qquad (1)$$

which is content-dependent, because the signal terms *x* and *f(x)* survive in the difference of distortion functionals. For content-independence, the above difference should be some function of only the distortion term ε and should not contain *x* or any of signal derived from it.

Now we take a different route and take as a reference a unique image y. We then measure the distortions between x and $x+\varepsilon$, using y and $y+\eta$ as reference images, $y+\eta$ represents the doctored version of the reference image. The relationship of these signals and the distortion vis-à-vis the reference images y and $y+\eta$ is illustrated in Fig. 1. In this figure, the length of the vector $\vec{x}\,\vec{y}$ is simply equal to $M(x,y)$. The distance between the 1s of the vectors $\vec{x}\,\vec{y}$ and $\vec{x}(y+\eta)$ is $d = M(x,y) - M(x,y+\eta)$, and similarly $d' = M(x+\varepsilon, y) - M(x+\varepsilon, y+\eta)$ denotes the distance between the 1s of the dashed pair of vectors. For the case of the mean-square distortion it follows that:

$$d = E[(y-x)^2 - (y-x)^2 +$$
$$2(y-x)\eta - \eta^2] = E[2(y-x)\eta - \eta^2]$$

and

$$d' = E[(x+\varepsilon-y)^2 - (x+\varepsilon-y)^2 + 2(x+\varepsilon-y)\eta - \eta^2]$$
$$= E[2(y-x)\eta + 2\eta^*\varepsilon - \eta^2].$$

Now if one considers the difference of d and of d' one can observe that one achieves content-independence, that is:

$$D_1 \stackrel{\Delta}{=} d' - d = 2E[\eta^* \varepsilon] \qquad (2)$$

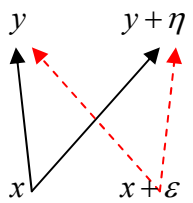


Fig. 1: Configuration of the signal vectors: the original image *x*, its tampered version $x+\varepsilon$, the reference image *y* and its tampered version $y+\eta$.

Let's consider another measure, the correlation coefficient, given by $M(a,b) \stackrel{\Delta}{=} E[ab]$. One can easily show that:

$$d = E[xy] - E[x(y+\eta)] = -E[x\eta] \text{ and}$$
$$d' = E[(x+\varepsilon)y] - E[(x+\varepsilon)(y+\eta)] = -E[x\eta] - E[\varepsilon^*\eta)]$$

so that $D_2 \stackrel{\Delta}{=} d' - d = -E[\varepsilon^* n]$. Again the difference of distortions is not a function of image content, *x* and *y*, but only of the product of distortions, $\varepsilon^* n$.

We can show that this property holds more generally if the second and higher order partials of the $M(x,y)$ functional are independent of *x* and *y*. Consider a generic D :

$$D = M(x,y) - M(x,y+\eta) - M(x+\varepsilon,y) + M(x+\varepsilon,y+\eta)$$

and its variational differential

$$\delta D = -M_x(x,y)\delta x - M_y(x,y)\delta y + M_x(x,y)\delta x + M_y(x,y)\delta y + M_{xy}(x,y)\delta x\delta y \ldots$$

where $M_{x^k y^m}(x,y) = \dfrac{\partial^{k+m} M(x,y)}{\partial x^k \partial y^m}$ [3]. This expression becomes:

$$\delta D = M_{xy}(x,y)\delta x\delta y + high\ \ order\ \ terms \ldots \qquad (3)$$

If the higher order partials of $M(x,y)$ are constant (or zero, as in the cases of the mean-square distortion and correlation coefficient), then the content-independence condition holds.

3. Experimental Results

We selected four measures from the list of image quality measures presented in [1], using Sequential Floating Forward Search (SFFS) algorithm. These three measures, as detailed in the Appendix were the two first-order moments of the angular correlation and two first-order moments of the Czenakowski measure.

We then used a training set of original images and their processed versions, as well as, the original and processed versions of the reference images. We used randomly selected reference images. A linear regression classifier was then designed using the statistics collected with the database of images [3].

The image alterations we experimented with were scaling, rotation, brightness adjustment and contrast enhancement. We trained and tested classifiers for brightness adjustment and contrast enhancement operations separately. In addition, we considered a mixture of alterations, which included scaling, rotation, brightness and contrast enhancements, and designed a classifier for mixed sequential processing. An image database was formed by selecting images from [4] in order to carry out the simulations. The database in [4] contains a rich variety of 2000 images, from which 200 were chosen randomly. Half of the images were used in the training and the remaining in testing.

Table I: The performance of the classifiers

Image Alteration Type	False Positive	False Negative	Accuracy
Brightness Adj.	0/100	23/100	88.5%
Contrast Adj.	6/100	30/100	82%
Mixed Proc.	5/100	12/100	91.5%

The classification accuracies of the detectors designed for specific operations are given in Table I. In these experiments, the entire image was subjected to the same type of operation, as listed in the first column of Table I.

To illustrate how well the selected features capture the impact of the signal processing operations and how well they separate into clusters, we show scatter plots for brightness adjustment, contrast enhancement and mixed sequential processing in Figures 1 a, b and c, respectively. In these figures the axes represent a subset of three features out of the four used. Each figure displays the scattering of the three features obtained from 200 unprocessed (blue), 200 processed (red) images. The axis denoted by d1 and d2 are the standard deviations of angular correlation measure and Czekanowski similarity measures respectively. Third axis d3 is the standard deviation of another correlation measure.

In a second set of more realistic experiments, we addressed the testing of "doctored images". We doctored 16 images by either inserting extra content or replacing the original content. To make them look like natural and avoid any suspicion, the inserted content had to be resized, rotated and brightness adjusted skillfully before pasting it to the image. In some cases we had to blur the block boundaries after pasting. While resizing and rotation were used in every doctored image, we had to do brightness adjustment only in a couple of images. We also obtained 44 doctored images from Internet. We tested 60 doctored images against brightness adjustment, contrast enhancement and mixed sequential processing classifiers. The results of the tests are given in Table II.

Table II: Performance of the classifiers

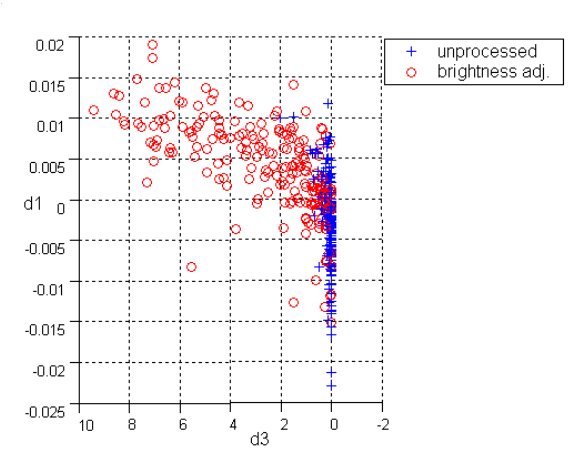
Image Alteration Type	False Positive	False Negative	Accuracy
Brightness Adj.	31/60	3/60	69.2%
Contrast Adj.	25/60	6/60	74.2%
Mixed Proc.	7/60	17/60	80.0%

4. Conclusions

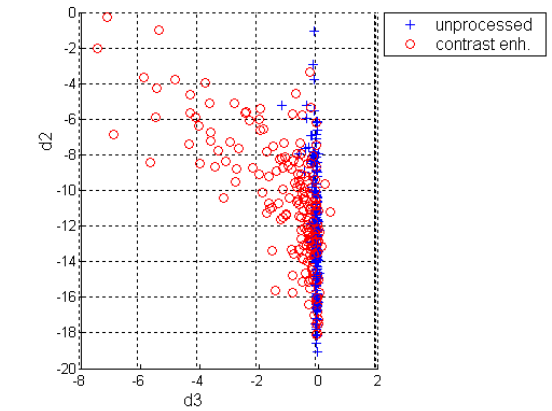
In this paper we proposed a framework for digital image forensics. First, we presented a novel way of content-independent distortion measurement within the framework of image forensics. Second, content-independent distortion measurements were used as features in the design of classifiers. The performance results were encouraging as we were able to discriminate a doctored image from its originals with a reasonable accuracy.

There is significant amount of work that still needs to be done. We need to perform more extensive testing of our classifier. The doctored images we used had a manipulated block sizes that were at least a 100 pixel wide. We need to create test data with smaller manipulations. Also, we need a data set of high quality manipulations as opposed to the ones we generated just for preliminary testing.

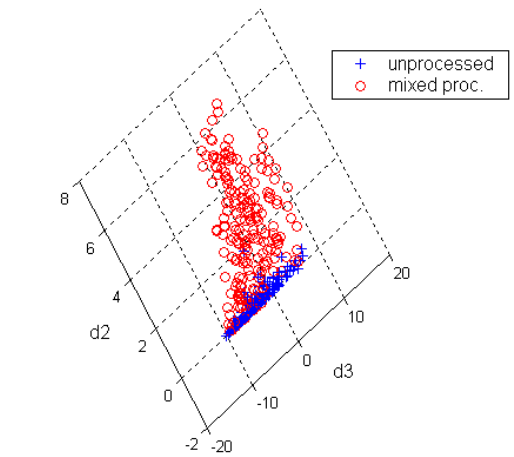
We are also investigating a larger variety of features and the use a more sophisticated classifier as compared to the simple linear classifier we use here.



a)



b)



c)

Figure 1. The scatter diagrams of features in a) brightness adjustment, b) contrast enhencement, c) mixed processing.

5. Appendix

The three different distortion measures used in the paper are shown below. We denote the color components of a three band color image at the pixel position i, j, and in band k as $C_k(i,j)$, where $k = 1,...,3$ and $i, j = 1,..., N$. The boldface symbols $\mathbf{C}(i,j)$ and $\hat{\mathbf{C}}(i,j)$ indicates the color pixel vectors, respectively, of the original and processed image. $\mathbf{C}$ itself denotes a color image. The norm and inner product of vectors are defined as $\|\mathbf{C}(i,j)\| = \sqrt{C_1(i,j)^2 + C_2(i,j)^2 + C_3(i,j)^2}$

$$\left\langle \mathbf{C}(i,j), \hat{\mathbf{C}}(i,j) \right\rangle = C_1(i,j)\hat{C}_1(i,j) + C_2(i,j)\hat{C}_2(i,j) + C_3(i,j)\hat{C}_3(i,j)$$

respectively.

First Order Statistics Of Angular Correlation Measure

$$\cos(\Theta_{ij}) = \frac{\left\langle \mathbf{C}(i,j), \hat{\mathbf{C}}(i,j) \right\rangle}{\|\mathbf{C}(i,j)\| \left\|\hat{\mathbf{C}}(i,j)\right\|}, \quad \mu_\theta = \frac{1}{N^2} \sum_{i,j=0}^{N-1} |\cos(\Theta_{ij})|,$$

$$d_1 = \left[\frac{1}{N^2} \sum_{\mathrm{i,j=0}}^{\mathrm{N-1}} \left(\cos(\Theta_{\mathrm{ij}}) - \mu_\theta\right)^2 \right]^{1/2}$$

First Order Statistics of Czekanowski Similarity Measure

$$\chi_{ij} = \frac{2\left\langle \mathbf{C}(i,j), \hat{\mathbf{C}}(i,j) \right\rangle}{\|\mathbf{C}(i,j)\| + \left\|\hat{\mathbf{C}}(i,j)\right\|}, \qquad \mu_\chi = \frac{1}{N^2} \sum_{i,j=0}^{N-1} |\chi_{ij}|,$$

$$d_2 = \left[\frac{1}{N^2} \sum_{\mathrm{i,j=0}}^{\mathrm{N-1}} \left(\chi_{\mathrm{ij}} - \mu_\chi\right)^2 \right]^{1/2}$$

$$\nu_{ij} = \frac{\|\mathbf{C}(i,j)\|}{2\left(\left\|\hat{\mathbf{C}}(i,j)\right\| + \left\langle \mathbf{C}(i,j), \hat{\mathbf{C}}(i,j) \right\rangle \right)},$$

$$\mu_\nu = \frac{1}{N^2} \sum_{i,j=0}^{N-1} |\nu_{ij}|, \; d_3 = \left[\frac{1}{N^2} \sum_{\mathrm{i,j=0}}^{\mathrm{N-1}} \left(\nu_{\mathrm{ij}} - \mu_\nu\right)^2 \right]^{1/2}$$

6. References

[1] I. Avcibas, N. Memon, B. Sankur, "Steganalysis Using Image Quality Metrics", *IEEE Trans. on Image Processing*, Vol. 12, pp. 221-229, February, 2003.

[2] I. Avcibas, B. Sankur, K. Sayood, "Statistical Evaluation of Image Quality Measures", *Journal of Electronic Imaging*, Vol. 11, pp. 206-223, April, 2002.

[3] A. C. Rencher, *Methods of Multivariate Analysis*, New York, John Wiley (1995).

[4] Image Steganography Database – Dartmouth University. http://www.cs.dartmouth.edu/~farid/.