ANTI-FORENSICS OF MEDIAN FILTERING

Zhung-Han Wu, Matthew C. Stamm, K. J. Ray Liu

Dept. of Electrical and Computer Engineering, University of Maryland, College Park

ABSTRACT

A number of forensic techniques have been developed to identify the use of digital multimedia editing operations. In response, several anti-forensic operations have been designed to fool forensic algorithms. One operation that has received considerable attention is median filtering, since it can be used for image enhancement or anti-forensic purposes. As a result, several median filtering detectors have been developed. In this paper, we propose an anti-forensic technique to disguise the use of median filtering. We do this by first proposing a model for an unaltered image's pixel difference distribution. We then modify a median filter image's pixel difference distribution using anti-forensic noise so that it no longer contains median filtering fingerprints. Through a series of experiments, we are able to show that our anti-forensic technique can fool existing median filtering detectors under realistic conditions.

Index Terms— Anti-Forensics, Median Filter, Pixel Difference

1. INTRODUCTION

Due to the ease with which multimedia content can be manipulated, measures to verify the authenticity of multimedia content are in pressing need. Several forensic measures on multimedia contents have been developed to identify different multimedia editing operations. For example, some techniques are designed to identify image resampling [1], double JPEG compression [2], contrast enhancement [3] and median filter operation [4], [5]. These forensic techniques operate by detecting the forensic fingerprints left in the multimedia content by editing operations.

An intelligent forger, however, can develop anti-forensic measures in order to fool the forensic detectors. These anti-forensic measures aim at eliminating the fingerprints introduced by multimedia editing operations. Several anti-forensic techniques have been developed for different operation detectors, such as removing the artifact of resizing and rotation in [6], forging the history of image compression [7], [8], histogram manipulation [9], color filter array pattern alteration [10], and video motion vector alteration for frame deletion in video sequence in [11]. By studying anti-forensics, we can identify the capabilities of an intelligent forger and develop measures to detect the use of anti-forensic techniques.

Median filtering is an image editing operation of particular forensic significance. It is commonly used to perform several image enhancement tasks such as noise suppression and smoothing. Median filtering has the useful property of preserving edge content due to its nonlinear nature. Recently, median filtering has been shown to be destructive to several other image manipulation traces due to its strong nonlinearity [6]. Due to this destructive nature of median filter, the median filter has been integrated into several anti-forensic techniques such as in [8]. Thus forensic measures on median filtering is essential because once median filter is detected by the detector, the authenticity of the image is questionable due to the possibility of other multimedia operation. Several median filter detectors have been proposed in [4] and [5] for detecting median filter operations.

An anti-forensic attack on median filtering detectors is of particular interest for image forgers. While median filtering is destructive to other image manipulation fingerprints, it leaves behind its own fingerprints. Consider the case that a forger applies a median filter at the end of image processing to destroy the fingerprints of previous image editing operations. The forger would then like to apply an anti-forensic technique to remove the median filter fingerprints from the image. If this can be done, the forger can produce a forged image that is free from any image editing operation fingerprints. This scenario shows that the destructive nature of median filtering makes median filtering anti-forensics very appealing to forgers.

In this paper, we propose a novel anti-forensic technique to remove traces of median filtering. We first propose using a generalized Gaussian distribution to model the pixel value difference distribution of an unaltered and median filtered image. We develop an estimation method to estimate a plausible pixel value difference distribution of an unaltered image given a median filtered one. Using the plausible estimated pixel value difference distribution as a target, we estimated the noise distribution to be added in the median filtered image and shift the pixel value difference distribution to the plausible one. Several measures are applied in the noise addition algorithm to balance the attack strength and the visual perception. Finally the anti-forensic attack is tested against several median filter detectors and the attack successfully falsifies the detector in typical detection scenarios. While prior work removes traces of median filter by optimally sharpening an image [12], our proposed apporach achieves better performance at low false alarm rates.

2. PROBLEM FORMULATION

We first give a brief introduction on the operation of median filter. Let $x_{i,j}$ be the pixel value at location (i, j) of an unaltered image. The corresponding pixel $y_{i,j}$ in a median filtered version image is given by $y_{i,j} = \text{med}_s(x_{i,j})$. Here med_s denoted the median filtering operation with a square filter window with size of s defined as

$$\text{med}_s(x_{i,j}) = \\ \text{median}\{x_{l,m} | 0 \leq \lfloor |i-l|/2 \rfloor \leq s, 0 \leq \lfloor |j-m|/2 \rfloor \leq s\}. \quad (1)$$

2.1. Median Filter Detectors

Several median filter detectors have been proposed in prior works [4], [5]. Kirchner and Fridrich proposed a detector that operates based on the statistics of pixel value differences of a median filtered image and an unaltered image [4]. The detector first divides the image into blocks, the pixel difference histogram in each of the block is calculated. Next, the ratio ϱ of the number of pixel differences

The authors can be contacted at {zhwu, mcstamm, kjrliu}@umd.edu

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

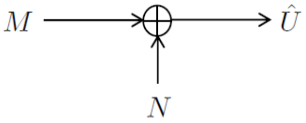

Fig. 1. System model for the anti-forensic technique.

whose value is zero to the number of pixel differences whose value is one is calculated for each block. A weighting function is applied to the ϱ value of each block to avoid statistical distortion in saturated regions and a final measure $\hat{\varrho}$ of the strength of median filtering fingerprints is obtained.

Kirchner and Fridrich also proposed using the subtractive pixel adjacency matrix (SPAM) detector which is originally developed in steganalysis in [13] for the detection of median filter operation. They first model the pixel value difference distributions in an image to be an nth order Markov chain. The SPAM detector then uses the Markov chain transition probabilities as features vectors which is used in support vector machine.

In [5], Yuan proposed a median filter detector which collects blockwise Median Forensic Features (MFF) which are statistics based on the pixel values and its distribution in the block. Different entries of MFF is then combined heuristically to produce a new index f. A binary decision is then made according to the index f to determine whether the image has undergone median filter or not.

2.2. Anti-Forensic Technique

Both the ϱ and SPAM detectors proposed by Kirchner and Fridrich detect median filtering by capturing features of the distribution of an images pixel value differences. Furthermore, several of the detection features proposed by Yuan implicitly capture information about an images pixel difference distribution. As a result, an anti-forensic attack on median filter detector should ensure that the pixel difference distribution of an anti-forensically modified image resembles one that is from an unaltered image.

Since a forger does not know whether a median filtering detection technique will examine the horizontal or vertical pixel differences, they must insure that the distribution in both directions contain no evidence of median filtering. Additionally, since pixel values are correlated in both the horizontal and vertical directions, a forger needs to ensure that modifications to the pixel differences in one direction do not create visual distortions in the other direction. As a result, we examine the pixel differences jointly in both directions. We define the pixel value difference pair $d_{i,j}$ as

$$d_{i,j} = \begin{pmatrix} x_{i,j} - x_{i,j+1} \\ x_{i,j} - x_{i+1,j} \end{pmatrix} = \begin{pmatrix} h_{i,j} \\ v_{i,j} \end{pmatrix}. \qquad (2)$$

Our goal is to fool median filtering detectors by modifying the pixel differences in a median filtered image so that its pixel difference distribution appears to come from an unaltered image. We do this by first parametrically modeling the pixel value difference distribution of an unaltered image and a median filtered image. We modify the pixel difference distribution of a median filtered image by adding specially designed noise its pixel differences. We design the distribution of this two dimensional noise so that when it is added to the pixel differences in a median filtered image, the pixel difference distribution of the anti-forensically modified image will match an estimate of the images pixel difference distribution before median filtering.

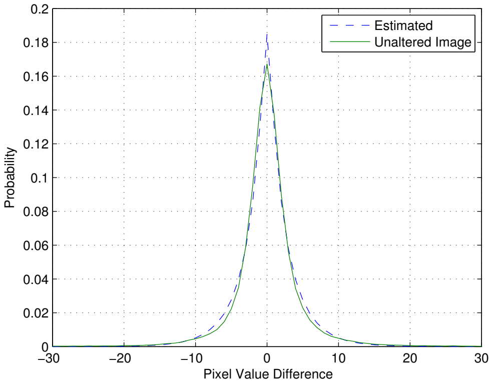

Fig. 2. Comparison of generalized Gaussian and actual histogram of pixel value difference in an unaltered image.

Let the random variables U and M be the pixel value difference of an unaltered image and a median filtered image, respectively, and N be a random variable representing the anti-forensic noise. Then by exploiting the property that when two random variables are added together, their distributions convolve with each other

$$f_U(d_{i,j}) = f_M(d_{i,j}) * f_N(d_{i,j}), \qquad (3)$$

where f_M, f_N and f_U are the distributions of M, N and U respectively. We can see from Eq.(3) that designing the anti-forensic noise distribution is equivalent to the design of a system impulse response $f_N(d_{i,j})$ with predefined input $f_M(d_{i,j})$ and output $f_U(d_{i,j})$. The system overview is illustrated in Fig. 1

We approach this task by the following steps. We first propose using generalized Gaussian distribution to characterize the distribution of pixel value difference vector d. The parameters for the generalized Gaussian distribution for both unaltered $f_U(d_{i,j})$ and median filtered $f_M(d_{i,j})$ images are gathered using an image database, and thus we find the relationship of the distribution parameters between the median filtered image and an unaltered one. We use the relationship to estimate a plausible unaltered image distribution $\hat{f}_U(d_{i,j})$ given a median filtered one. By exploiting the relation in Eq. 3, we can design the noise distribution $f_N(d_{i,j})$ to add to the image. We devise technique to add the noise realization into the image. Finally our technique is tested against the known detectors to verify the anti-forensic attack performance.

3. TARGET DISTRIBUTION ESTIMATION

3.1. Parametric Model of Pixel Value Difference Distribution

We propose to model the pixel value difference distribution using a generalized Gaussian distribution. Fig.2 and Fig.3 shows that the generalized Gaussian fits the pixel value difference distribution. Since we care about the joint distribution of the horizontal and the vertical pixel differences, we further propose to model the pixel value difference distribution using a two dimensional generalized Gaussian distribution. The formula for a two-dimensional generalized Gaussian distribution is given in Eq. (4) [14],

$$f_d(d;\alpha,\Sigma) = \frac{\det \Sigma^{-1/2}}{(Z(\alpha)A(\alpha))^2} \exp\left(- \left|\left| \frac{\Sigma^{-1/2} d}{A(\alpha)} \right|\right|_\alpha^\alpha \right), \qquad (4)$$

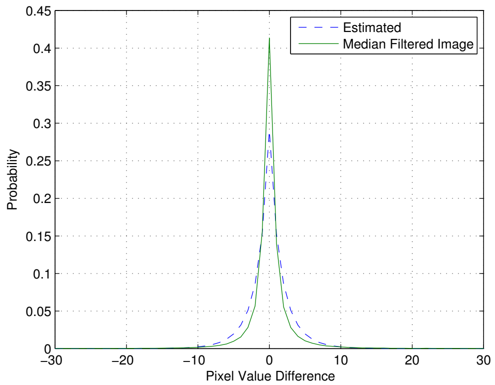

Fig. 3. Comparison of generalized Gaussian and actual histogram of pixel value difference in a median filtered image.

where $Z(\alpha)$ and $A(\alpha)$ are normalizing constants.

In Eq.(4), d is a 2 by 1 vector. Σ is the covariance matrix of the random vector and α is a parameter which controls the shape of the generalized Gaussian distribution. The distribution can have different kurtosis by adjusting the parameter α.

3.2. Parametric Estimation of a Plausible Unaltered Image Pixel Value Difference Distribution

We make an assumption that the pixel value difference distribution is directionally invariant. In order the parameterize the distribution, we need to estimate Σ and α. We obtain an estimate of the covariance matrix $\hat{\Sigma}$ and the maximum likelihood estimator for $\hat{\alpha}$ using the method proposed in [14]. In order to increase the stability of the estimation algorithm, we apply a Gaussian filter on the histogram of the image before the estimation.

Next, we estimate the parameters Σ^U and α^U in the unaltered images pixel difference distribution directly from the median filtered images parameters Σ^M and α^M. We do this by estimating Σ^U, α^U, Σ^M, and α^M from each image in a training database, then we perform a linear regression on these parameters using a least squares fit. For example

$$\alpha^U = a_0 + a_1\Sigma^M_{1,1} + a_2\Sigma^M_{1,2} + a_3\Sigma^M_{2,1} + a_4\Sigma^M_{2,2} + a_5\alpha^M. \tag{5}$$

Using these estimated parameters and our parametric model, we can obtain an estimate of the unaltered images pixel difference distribution $\hat{f}_U$.

Now we have the parameters for f_M and $\hat{f}_U$ and we can design a distribution f_N using the relation Eq. (3). We perform the estimation in the Fourier transform domain

$$F_N(\omega) = \hat{F}_U(\omega)/F_M(\omega). \tag{6}$$

where $F(\omega) = DFT\{f(d)\}$. Then f_N can be obtained by using IDFT on $F_N(\omega)$. f_N would have small portion of negative parts because this is not a system impulse response but a probability distribution which is strictly non-negative. We project it to the closest set of probability distribution by truncating the negative parts and normalizing the distribution.

4. NOISE ATTACK ALGORITHM

Since the anti-forensic methodology is to add noise to the pixel differences $d_{i,j}$, the pixel values must be recovered from the pixel differences. This means that when reconstructing the pixel values from the pixel differences, we need to value of at least one reference pixel. We refer to this pixel as the anchor point. Assuming that we choose the pixel at the location (i, j) to be the anchor point, the corresponding pixel value $y_{i,j}$ in the anti-forensically modified image is given by $y_{i,j} = x_{i,j}$.

Once the anchor point is chosen, we then modify all the pixels in the i^{th} row using

$$y_{i,j+l} = x_{i,j} - \sum_{k=0}^{l-1} h_{i,j+k} - \sum_{k=0}^{l-1} n^h_{i,j+k}. \tag{7}$$

where $n^h_{i,j+k}$ is the noise realization of the obtained from the one dimensional noise distribution f_N using the acceptance and rejection method.

Next, we modify the rows below and above the i^{th} row using

$$y_{i+k+1,j+l} = y_{i+k,j+l} - v_{i+k,j+l} - n^v_{i+k,j+l}. \tag{8}$$

where the probability distribution of $n^v_{i+k,j+l}$ is given by the conditional distribution $f_N(n^v_{i+k,j+l}|n^h_{i+k,j+l} = y_{i+k,j+l} - y_{i+k,j+l+1})$. This conditional distribution can be obtained from the joint distribution $f_N(d_{i,j})$. We continue the process until all rows are modified accordingly.

4.1. Distortion Limiting Measures

Though anti-forensically modifying an image in this manner will remove median filtering fingerprints from an image, it will also introduce distortion. If a large noise value is added to $d^M_{i,j}$, then the reconstructed adjacent pixel value $y_{i,j+1}$ may have a large deviation from its original value $x_{i,j+1}$. This deviation would further propagate to the next adjacent pixel $x_{i,j+2}$ because the reconstruction depends on the value of $x_{i,j+1}$. As a result, the further that we move from our anchor point, the greater the possible distortion introduced into the image. To mitigate the distorting effects of anti-forensics, we add the following modifications to our algorithm:

We partition the image into blocks and select the anchor point for reconstruction to be at the center of the block. Since the potential amount of distortion increases the further we move from our anchor point, we segment the image into blocks and independently anti-forensically modify each block. This prevents each the modified pixel values from drifting too far from their original values. By doing this, the horizontal and vertical distance between any point in the block and the anchor point is no greater than half of the block width.

Additionally, since the distortion incurred by a single large noise value can propagate to a large area, we limit the range of values that the anti-forensic noise can take to $[-T, T]$. The block partition and the anchor point selection can not prevent the possible large noise realization value which cause large distortion, by selecting proper T, we can reduce the distortion in a controlled fashion.

While these measures help decrease the visual distortion introduced into an image by anti-forensics, they negatively impact the effectiveness of our anti-forensic technique. We have experimentally observed that limiting the range of noise values $[-T, T]$ reduces the effect of the anti-forensic noise. To compensate for this, we multiply the covariance matrix of the target distribution by a correction factor $\beta < 1$. This slightly increases the variance of the anti-forensic noise and helps overcome the negative effects of limiting the range of the noise values.

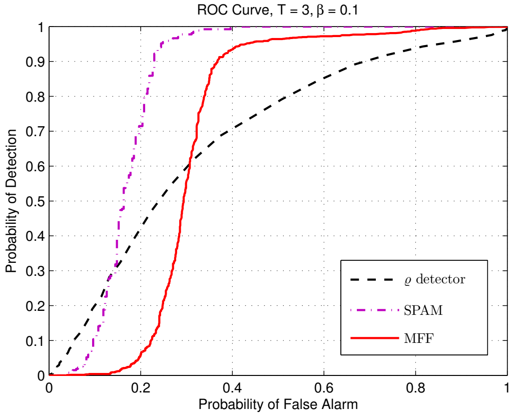

Fig. 4. ROC curves for three detectors with $T = 3$ and $\beta = 0.1$

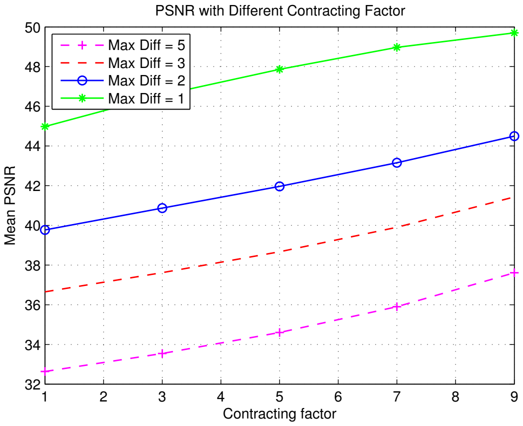

Fig. 5. PSNR curve for different T and β

5. SIMULATION RESULT

We evaluated the performance of our anti-forensic technique using the UCID [15] database which consists of 1338 color images which had never been compressed. All the images were first converted to gray scale image before any further processing. The gray scale image was used directly as the unaltered image. For the median filtered image database, the grayscale images were processed using a median filter with support 3.

To measure the baseline performance of the ϱ, SPAM, and MFF detectors, we used each forensic technique to test for median filtering in both our databases of unaltered and median filtered images. From these detection results, we found that all three detectors were able to achieve perfect detection, i.e. each achieved a probability of detection of $P_D = 100\%$ with a corresponding probability of false alarm of $P_{FA} = 0\%$.

Next, we produced our anti-forensically modified image database by selecting different parameters for the anti-forensic attack algorithm and produced corresponding sets of modified image databases. The noise realization limit T was varied between $1, 2, 3, 5$, and the correction factor β was selected to be $0.1, 0.3, 0.5, 0.7, 0.9$. Then the performance of the anti-forensic technique was tested against the three detectors. The PSNR between the median filtered images and the anti-forensically modified images was also measured to evaluate how much distortion was added during the anti-forensic modification.

In typical forensic settings, there is a high cost associated with false alarms. For example, in legal scenarios it is unlikely that a forensic finding will be allowed into evidence if there is a significant probability that a forgery detection corresponds to a false alarm. As a result, a forensic investigator must typically operate with a false alarm rate less than 20%. From a forgers point of view, it is critical that any anti-forensic countermeasures they use significantly reduce the performance of forensic techniques in this false alarm region. Since a forensic investigator cannot typically perform detection outside of this critical region, the performance of forensic techniques at higher false alarm rates is of less concern.

The ROC curves for the ϱ, SPAM, and MFF detectors are given in Fig.4 with $T = 3$ and $\beta = 0.1$. The detection rate for the three detectors after the anti-forensic modification is mostly less than 50% in the critical false alarm rate region. Specifically, for $P_{FA} < 15\%$, P_D is typically under 50%. This shows that our anti-forensic technique is very effective in the critical false alarm region. Also, Although our anti-forensic technique alters the pixel value difference distribution, our technique also works for the MFF detector that does not rely directly on the pixel value difference distribution.

The PSNR between the median filtered image and the anti-forensic modified image with different T and β is given in Fig.5. With a given T, PSNR decreases with smaller β; while with a given β, the PSNR decreases with larger T. The larger T and a smaller β mean that the stronger the modification is. As we can see, most of the parameter combination would result in PSNR in the high 30s. Note that the PSNRs are provided as a reference to show that the images are of sufficient quality. In reality, a forensic investigator does not have access to the original image and cant evaluate the quality loss. All they can do is make a subjective judgment of whether the image appears to be authentic or not. The tradeoff between the anti-forensic technique parameters and the visual quality means we can optimize between the visual quality and the attack strength to ensure the attack while preserving image qualities.

6. CONCLUSION

In this paper, we proposed an anti-forensic technique to fool forensic median filtering detectors that operates by adding anti-forensic noise to an images pixel difference distribution. To accomplish this, we proposed using a two dimensional generalized Gaussian distribution to model an images pixel value difference distribution. We estimated the distribution parameters for both unaltered images and median filtered images, then used linear regression to estimate an unaltered distribution parameters from a median filtered images parameters. We design our anti-forensic noises distribution so that the pixel difference distribution of an anti-forensically modified image appears to come from an unaltered image. In order to ensure the visual quality of the anti-forensic modification, we proposed several measures for limiting the distortions introduced in the modification. Finally, our anti-forensic technique was tested against several median filter detectors. The results indicate that our anti-forensic technique can fool existing median filtering detectors under realistic scenarios.

7. REFERENCES

[1] Alin C. Popescu and Hany Farid, "Exposing digital forgeries by detecting traces of re-sampling," *IEEE Trans. Signal Process.*, vol. 53, pp. 758–767, 2004.

[2] T. Pevný and J. Fridrich, “Detection of double-vompression in jpeg images for applications in steganography,” *IEEE Trans. Inf. Forensics and Security*, vol. 3, no. 2, pp. 247 –258, Jun. 2008.

[3] M.C. Stamm and K.J.R. Liu, “Forensic detection of image manipulation using statistical intrinsic fingerprints,” *IEEE Trans. Inf. Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010.

[4] M. Kirchner and J. Fridrich, “On detection of median filtering in digital images,” in *Media Forensics and Security II, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 18-20, 2010, Proceedings*, Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp, Eds. 2010, vol. 7541 of *SPIE Proceedings*, p. 754110, SPIE.

[5] H.-D. Yuan, “Blind forensics of median filtering in digital images,” *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1335–1345, Dec. 2011.

[6] M. Kirchner and R Böhme, “Hiding traces of resampling in digital images,” *IEEE Trans. Inf. Forensics and Security*, vol. 3, no. 4, pp. 582–592, Dec. 2008.

[7] M.C. Stamm, S.K. Tjoa, W.S. Lin, and K.J.R. Liu, “Anti-forensics of JPEG compression,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, Mar. 2010, pp. 1694 –1697.

[8] M.C. Stamm and K.J.R. Liu, “Anti-forensics of digital image compression,” *IEEE Trans. Inf. Forensics and Security*, vol. 6, no. 3, pp. 1050–1065, Sept. 2011.

[9] M. Barni, M. Fontani, and B. Tondi, “A universal technique to hide traces of histogram-based image manipulations,” in *Proceedings of the on Multimedia and security*, New York, NY, USA, 2012, MM&#38;Sec ’12, pp. 97–104, ACM.

[10] M. Kirchner and R Böhme, “Synthesis of color filter array pattern in digital images.,” in *Media Forensics and Security*, Edward J. Delp, Jana Dittmann, Nasir D. Memon, and Ping Wah Wong, Eds. 2009, vol. 7254 of *SPIE Proceedings*, p. 72540, SPIE.

[11] M.C. Stamm, W.S. Lin, and K.J.R. Liu, “Temporal forensics and anti-forensics for motion compensated video,” *IEEE Trans. Inf. Forensics and Security*, vol. 7, no. 4, pp. 1315 – 1329, Aug. 2012.

[12] M. Fontani and M. Barni, “Hiding traces of median filtering in digital images,” in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug., pp. 1239–1243.

[13] T. Pevný, P. Bas, and J. Fridrich, “Steganalysis by subtractive pixel adjacency matrix,” in *Proceedings of the 11th ACM workshop on Multimedia and security*, New York, NY, USA, 2009, MM&#38;Sec ’09, pp. 75–84, ACM.

[14] L. Boubchir and J.M. Fadili, “Multivariate statistical modeling of images with the curvelet transform,” in *Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on*, 28-31, 2005, vol. 2, pp. 747 – 750.

[15] G. Schaefer and M Stich, “UCID - an uncompressed colour image database,” *Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia 2004*, pp. 472–480, 2004.