Image forgery detection by using No-Reference quality metrics

F. Battisti, M. Carli, A. Neri

Applied Electronics Department
Universitá degli Studi Roma TRE
Rome, Italy

Keywords: Image Forgeries, Digital Forensics, Cut-and-Paste Forgery detection, Image Quality metrics

1. ABSTRACT

In this paper a methodology for digital image forgery detection by means of an unconventional use of image quality assessment is addressed. In particular, the presence of differences in quality degradations impairing the images is adopted to reveal the mixture of different source patches. The ratio behind this work is in the hypothesis that any image may be affected by artifacts, visible or not, caused by the processing steps: acquisition (i.e., lens distortion, acquisition sensors imperfections, analog to digital conversion, single sensor to color pattern interpolation), processing (i.e., quantization, storing, jpeg compression, sharpening, deblurring, enhancement), and rendering (i.e., image decoding, color/size adjustment). These defects are generally spatially localized and their strength strictly depends on the content. For these reasons they can be considered as a fingerprint of each digital image. The proposed approach relies on a combination of image quality assessment systems. The adopted no-reference metric does not require any information about the original image, thus allowing an efficient and stand-alone blind system for image forgery detection. The experimental results show the effectiveness of the proposed scheme.

2. INTRODUCTION

Image splicing is an image forgery mechanism realized by cutting and pasting portions of some images into another one. Nowadays this operation is a simple task due the spreading of easy-to-use and freely available software. Post-processing is usually not required for melting regions belonging to different sources and if accurately performed it is almost impossible to visually detect the different zones. In this context authenticity verification is a challenging problem especially if auxiliary information is not provided to the verification system.

Digital image forensics techniques may help in the identification of regions that have been altered and manipulated. The available image forgery techniques can be divided into active and passive ones according to the presence of extra information.

Active system are base on watermarking methodologies[1–4] for tampering detection. They are based on the insertion of some features extracted from the image into the image itself. During the authentication control, the detection of modification in the hidden data can be used for assessing modification of the original image. The drawback of watermarking based methods is in the required cooperation of the image content creation system for the extra information embedding. Therefore active approaches are difficult to be effective due to the amount of digital data already available on the Internet.

Passive approaches do not require extra information and they can be considered blind with respect to the original image. These approaches rely on the extraction of some features from the image under test and, based on pre-defined rule or statistical thresholds, make a decision.

Several techniques for forensic detection have been proposed so far. However, due to the problem complexity, the available methods are effective in detecting specific image manipulations. A methodology for detecting any image modification is still to be achieved. The best solutions at the moment is to combine many tools in order to mix different approaches for revealing the presence of forgeries. Inconsistencies in light conditions are used

Further author information: (Send correspondence to F. Battisti)
Federica Battisti: E-mail: federica.battisti@uniroma3.it

in[5,6] as evidence of tampering. Briefly, the lighting environment is modeled with a nine dimensional model consisting of a linear combination of spherical harmonics. The parameters are extracted from a (2D and 3D) model of a persons face and head.

The detection of contrast enhancement and histogram equalization processes are used in[7] , the differences in camera response function have been investigated in[8,9] , and phase congruency in the under test image is adopted in[10] . Other approaches are based on non uniformity in image pixel correlation and image edge statistics[11] , or sharpening and blurring modification detection as in[12,13] . A different solution is proposed in[14] in which the authors present a method for matching the fingerprint of the camera that has taken a shot to a set of camera fingerprints stored in a database.

In this paper, the forgery of digital images performed by means of image splicing is detected by analyzing the scores of image quality metrics. The rest of the paper is organized as follows. In Section 3 a brief overview of typical image artifacts and of the metrics adopted for their impact evaluation is performed. In Section 3.1 the key elements of the selected quality metrics and the use of such indicators in the proposed method is presented. Section 3.2 describes the technique that has been used for the localization of the area suspected of tampering. In Section 4 the experimental results validating the system are reported and, finally, in Section 5 concluding remarks and future work are presented.

3. IMAGE QUALITY ASSESSMENT

The proposed method is based on the analysis of the characteristics of tampered digital images obtained by cut-and-paste procedures. In particular, the method aims at revealing the different impairment fingerprints, caused by the acquisition process, the color interpolation system, and the performed processing (i.e., quantization, storing, compression).

Many studies have been performed for understanding and classifying the introduced artifacts and for evaluating their impact on the original signal and on the final user. Blocking, ringing, and blurriness are probably the most perceivable ones. In the following a brief overview of these artifacts is reported:

- blockiness distortion (also known as blocking) is a distortion of the image characterized by the appearance of the underlying block encoding structure. Blockiness is often caused by coarse quantization of the spatial frequency components during the encoding process;
- ringing distortions are artifacts that appear as spurious signals (*rings*) near sharp transitions of the image. Visually, they appear as ”rings” near the edges;
- blurriness is defined as a spatial details loss and a reduction in the sharpness of edges in moderate to high frequency regions of the image such as in roughly textured areas or around scene objects.

The quantitative evaluation of each artifact is not an easy task. Several metrics have been designed to this aim. An overall solution is still far to be reached since often more that one artifact is present in the image, the intensity is not uniform in the image area, and the methodology for performing the measurements is still an open issue. All the objective image quality assessment metrics can be classified according to the amount of original information needed during the quality evaluation.

- Full-Reference (FR) methods[15,16] require the access to the reference image, that is assumed to have perfect quality. In practice, FR methods may not be applicable since very often the original image is not available.
- Reduced-Reference (RR) quality metrics[17,18] exploit partial information about the original image. However they require a cooperation between the image originator and the content .
- No-Reference metrics (NR) do not require any side information regarding the original media. For this reason, this class of metrics is the most promising in the context of broadcast scenario, since the original images is not available to end users. Designing effective NR metrics is a big challenge. Although human observers can usually assess the quality of an image without using the reference, creating a metric implementing such a task is difficult and, most frequently, results in a loss of performance in comparison

to the FR approach. Most of the proposed NR metrics estimate annoyance by detecting and estimating the strength of commonly found artifact signals. Among them, the metrics by Wu et al. and Wang et al. estimate quality based on blockiness measurements[19, 20] , while the metric by Caviedes et al. takes into account measurements of five types of artifacts[21] .

In the field of forensics forgery detection, FR metrics are not useful since they require the availability of the original image. In this case, a bitwise comparison is sufficient for creating the map of image alterations. Even RR metrics are of difficult applicability in this field. In fact those metrics require a preliminary agreement with the content creator for measuring and attaching the required features to the content; furthermore those features should be resilient to any change in data format, transcoding, etc.. The most viable solution to our goal is the use of NR metrics.

3.1 Proposed method

In the designed system a blind evaluation of local presence of each artifact is performed by using state of the art blind quality assessment methods, and finally to highlight the area of possible forgery, a features fusion approach has been proposed. In this work, we adopt a no-reference quality metric designed by Wang et al. in[22] that considers three important artifacts in the quality assessment. Based on the results achievable with this method, in our system the three features (F), Blocking (B), Activity (A) and Zero-Crossing (Z), have been used to authenticate digital images based on the evidence that a tampered image presents inconsistencies in image quality evaluation. More in detail, blurring is mainly due to the loss of high frequency DCT coefficients, while the blocking effect occurs due to the discontinuity at block boundaries, caused by the block-based quantization performed in the JPEG compression standard. The first part of the metric is based on the blocking estimation evaluated as the average differences across block boundaries. Other artifacts are taken into account by considering the reduction of the activity of the signal. The activity is evaluated using two factors: the average absolute difference between in-block image samples and the zero-crossing rate. In the following the basics behind the artifact estimation systems are reported for purpose of clarity.

The Blocking (B) in an image x of size [M,N] is accounted as the average differences across the boundaries of blocks of size 8×8 as:

$$B_h = \frac{1}{M\left(\lfloor \frac{N}{8} \rfloor - 1\right) \sum\limits_{i=1}^{M} \sum\limits_{j=1}^{\lfloor \frac{N}{8} \rfloor - 1} \left| d_h\left(i, 8j\right) \right|} \tag{1}$$

where B_h is the blocking feature computed on blocks of size 8×8, d_h is obtained by differencing the signal along each horizontal line:

$$d_h(m,n) = x(m,n+1) - x(m,n), n \in [1, N-1], m \in [1, M-1] \tag{2}$$

As previously mentioned, blurring is mainly due to the loss of high frequency DCT coefficients, while the blocking effect occurs due to the discontinuity at block boundaries, caused by the block-based quantization performed in the JPEG compression standard.

Other artifacts are taken into account by considering the reduction of the activity of the signal. The activity is evaluated using two factors: the average absolute difference between in-block image samples and the Zero-Crossing (Z) rate:

$$A_h = \frac{1}{7}\left[\frac{8}{M(N-1)} \sum_{i=1}^{M} \sum_{j=1}^{N-1} |d_h(i,j)| - B_h\right] \tag{3}$$

where A_h is the activity feature computed on blocks of size 8×8. Finally the Zero Crossing is evaluated as:

$$Z_h = \frac{1}{M(N-2)} \sum_{i=1}^{M} \sum_{j=1}^{N-2} z_h(m,n), \tag{4}$$

where where Z_h is the zero crossing feature computed on blocks of size 8×8, and

$$z_h(m,n) = \begin{cases} 1 & \textit{horizontal } Z \textit{ at } d_h(m,n) \\ 0 & \textit{otherwise} \end{cases} \tag{5}$$

Prior to feature extraction the original color image is represented in the YC_bC_r color space. This color space can be obtained by the RGB components as follows:

$$\begin{array}{l} Y = 16 + (65.481 \cdot R' + 128.553 \cdot G' + 0.114 \cdot B') \\ C_b = 128 + (-37.797 \cdot R' - 74.203 \cdot G' + 112 \cdot B') \\ C_r = 128 + (112 \cdot R' - 93.786 \cdot G' - 18.214 \cdot B') \end{array} \tag{6}$$

The features are extracted for every color component of the YC_bC_r image representation and then they are combined according to the following procedure:

1. the features obtained are first normalized in the range [0,1] and then linearly combined:

$$Sum = \sum_{i=Y}^{Cr} B_i + A_i + Z_i \; where \; i = Y, C_b, C_r; \tag{7}$$

2. in order to better exploit the information retrieved from the features, the correlation, c_j with $j = 1, .., 9$, between each metric and *Sum* is computed

3. the correlation values that have been obtained are used as weights in a new linear combination $Sum_{weighted}$:

$$Sum_{weighted} = \sum_{j=1}^{9} c_j F_j; \tag{8}$$

where F_j are the extracted features.

This procedure is performed to create a first localization map for identifying areas of possible tampering.

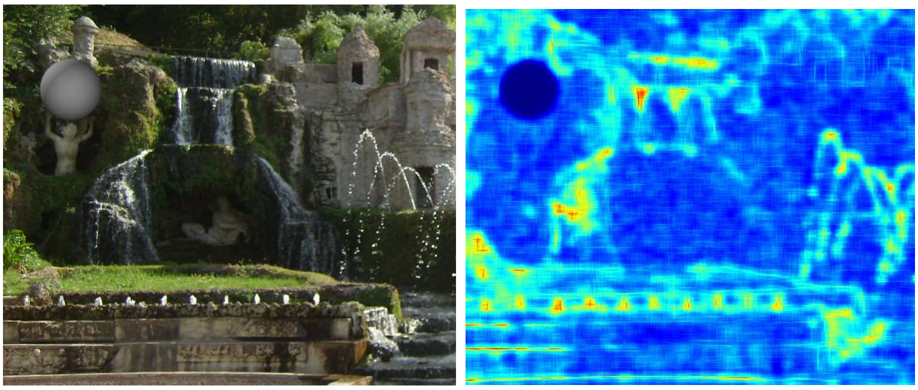
(a) (b)

Figure 1. Tampered image (a) and the corresponding first combination map (b).

3.2 Localization of suspected areas

In order to refine the map obtained after filtering, a post-processing is applied to the map.

To this aim, all the elements in the first combination map are classified in two groups: G_1, containing the information relative to the areas identified as *non modified*, and G_2 that contains the information on the areas recognized as *tampered*. In order to perform this grouping, each pixel in the first localization map is compared to a predefined threshold, T, as described in the following:

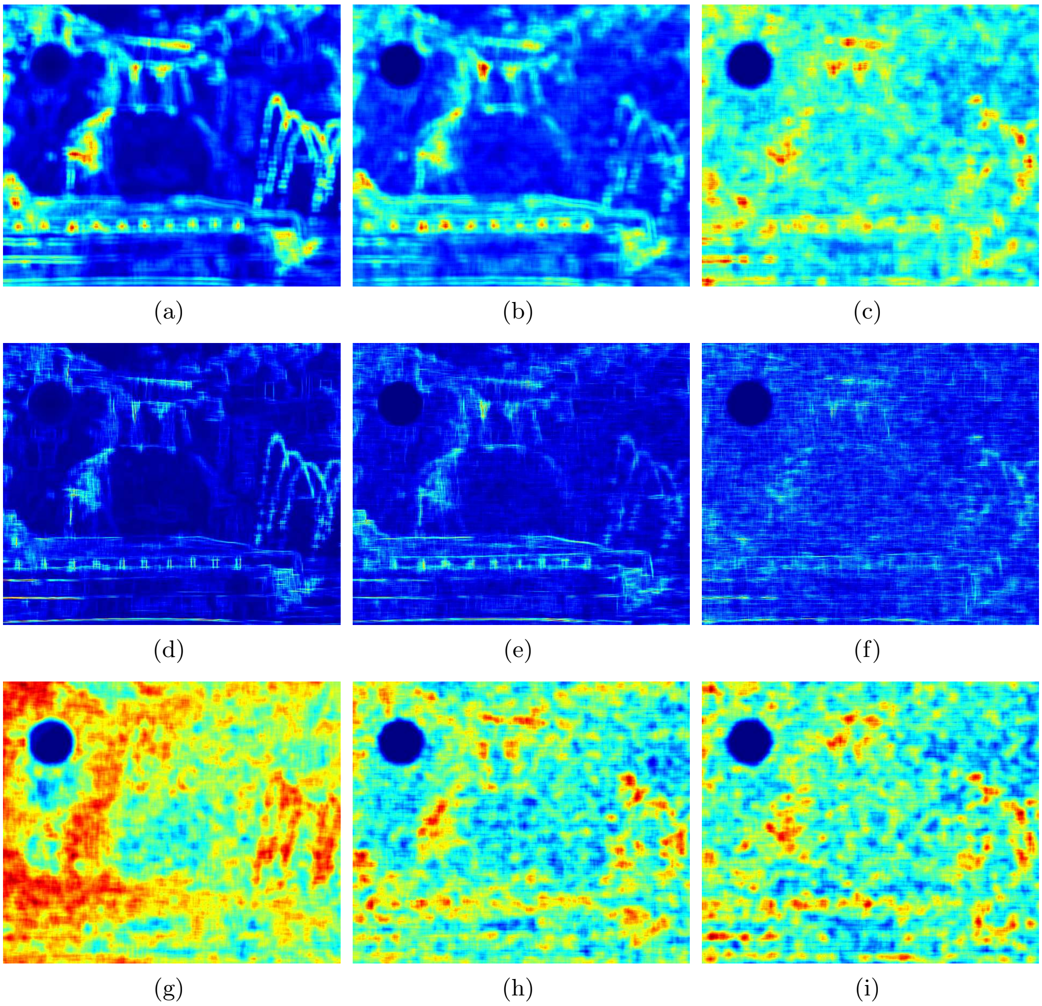

Figure 2. Y Cb Cr components of the *Activity* map of Figure 1, Y Cb Cr components of the *Blocking* map of Figure 1, and Y Cb Cr components of the *Zero-Crossing* map of Figure 1.

1. an initial estimate for T is selected. This value is set as the average grey level of the image;
2. the threshold is used to group the values in the map, M, in G_1 and G_2, according to their position (i, j)

$$\begin{cases} G_1 & if \quad \mathrm{M\,(i,j)} > T \\ G_2 & if \quad \mathrm{M\,(i,j)} < T \end{cases} \tag{9}$$

3. compute the average grey level values μ_1 and μ_2 for the pixels in regions G_1 and G_2
4. update the threshold value as:

$$T = \frac{\mu_1 + \mu_2}{2} \tag{10}$$

5. steps 2 through 4 are repeated until the variation of T, ΔT, in two successive iterations is smaller than a predefined parameter T_0.

When G_1 and G_2 are found, an edge detector is applied to the global threshold computation in order to emphasize the areas labeled as tampered in the final combination map (as shown in Figure 3).

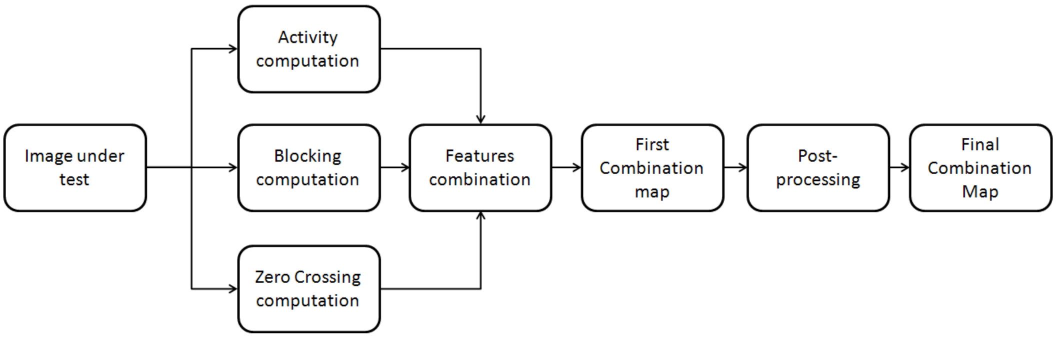

Figure 3. Block diagram of the proposed scheme.

4. EXPERIMENTAL RESULTS

To verify the effectiveness of the proposed method, it has been tested on two image tampered databases:

1. Database 1: it has been created by inserting cartoons in 26 color images by using Adobe Photoshop and Corel PaintShop Pro. The size of the images in this database ranges from 757 × 568 to 1152 × 768 pixels;
2. Database 2: Columbia Uncompressed Image Splicing Detection Evaluation Dataset*. This database contains 183 original images, and 180 spliced ones. The image sizes range from 757 × 568 to 1152 × 768 pixels and are uncompressed, in either TIFF or BMP formats. The spliced images are created using the original images, without any post-processing.

Figure 2 shows the *Activity*, *Blocking*, and *Zero-Crossing* indicators obtained from the analysis of the three color components Y, Cb, and Cr. As can be noticed, the presence of each artifact is detected with different intensity in each image component. Moreover, different metrics highlight the presence of artifacts in different areas of the image. To obtain an overall information of the possibly tampered areas, an optimization process has to be performed for combining the collected information. Figure 4 shows the results obtained for an image extracted from Database 2. Also in this case the tampered area is localized.

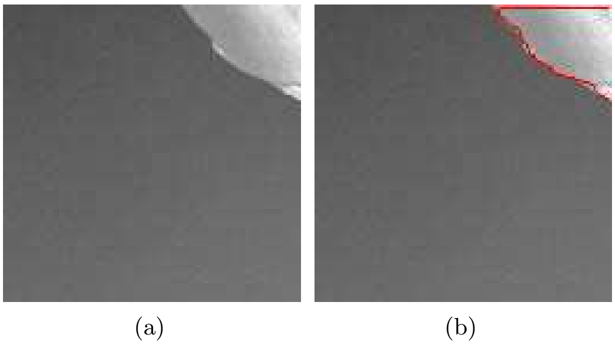
(a) (b)

Figure 4. Tampered image (a) and area detected as tampered (b).

Another example for an image taken from Database 2 is presented in Figure 6. The tampered area is identified with a red border. The experimental results, performed on the 2 databases, demonstrate the effectiveness of the proposed idea. The detection of the tampered area is successful in most of the tested images. This proves that image quality can be used for the detection of tampered areas. The fusion of information gathered by no-reference quality metrics can be useful in the forensics field to detect image anomalies. The accuracy in determining the borders of the tampered portions in the image can be limited by the presence of similar artifacts in the surrounding area. As can be noticed in Figure 7 (b), the final combination map presents high

[*] http://www.ee.columbia.edu/ln/dvmm/downloads/authsplcuncmp/

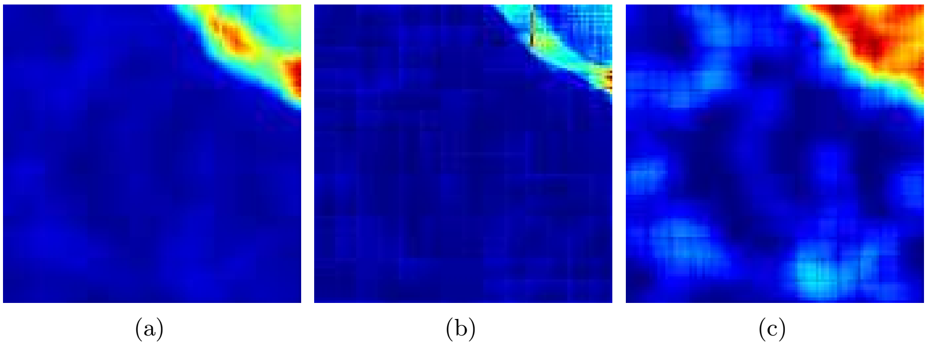
Figure 5. *Activity* map (a), *Blocking* map (b), and *Zero-Crossing* map (c).

values in positions corresponding to the two added cartoons. However, similar values are also in the left back part of the car. To further verify the effectiveness of the system, we tested for tampering a set of non tampered images. Even if distortions were detected by single metrics, the tampering indicator values resulting after the fusion step were below the alarm threshold.

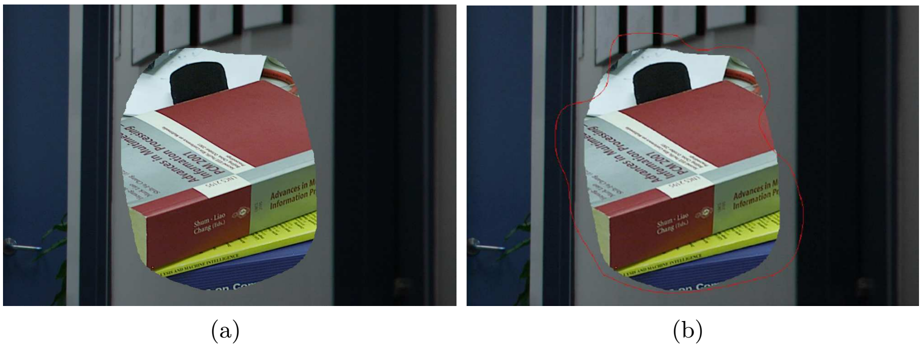
Figure 6. Tampered image (a) and area detected as tampered (b).

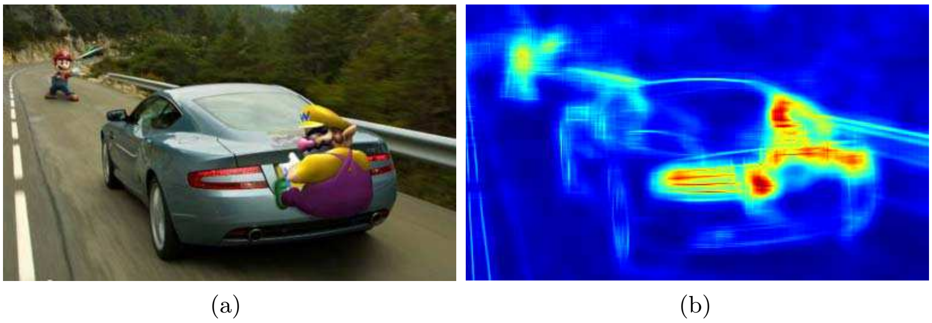
Figure 7. Tampered image (a) and final combination map (b).

5. CONCLUDING REMARKS

In this contribution we have demonstrated that the use of no-reference quality metrics can be adopted for recognizing tampered area in digital images. This method can be considered as a preliminary tool for inconsistencies evaluation. The proposed system is built on a combination of image quality degradation assessment systems. The adopted no-reference metric does not require any information about the original image, thus allowing an efficient and blind system for image forgery detection. The experimental results demonstrate the effectiveness of the proposed scheme. The analysis of the achieved results show a strong interaction between the performances of the selected quality metrics and the tampered area localization. Based on these considerations, work in progress is devoted to extending the considered artifacts by selecting a larger number of state of the art quality metrics. Furthermore, in order to improve the localization of tampered areas, we are focusing our efforts in the optimization of the fusion block.

REFERENCES

[1] Cox, I., Miller, M., and Bloom, J., [*Digital watermarking*], Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002).

[2] Barni, M. and Bartolini, F., [*Watermarking systems engineering. Enabling digital assets security and other applications*], no. ISBN 0824748069 in Signal Processing and Communications Series, CRC Press; 1 edition, New York (2004).

[3] Dittmann, J., "Content-fragile watermarking for image authentication," in [*Security and Watermarking of Multimedia Contents III*], Wong, P. W. and Delp, E. J., eds., *SPIE* **4314**, 175–184 (August 2001).

[4] Kung, C., Juan, K., Tu, Y., and Kung, C. in [*A Robust Watermarking and Image Authentication Technique on Block Property*], *Information Science and Engineering, 2008. ISISE '08. International Symposium on* **1**, 173–177 (December 2008).

[5] Johnson, M. and Farid, H., "Exposing digital forgeries in complex lighting environments," in [*Information Forensics and Security, IEEE Transactions on*], **2**, 450–461 (2007).

[6] Kee, E. and Farid, H. in [*Exposing Digital Forgeries from 3-D Lighting Environments*], *IEEE International Workshop on Information Forensics and Security* (December 2010).

[7] Stamm, M. and Liu, K. in [*Blind forensics of contrast enhancement in digital images*], *Image Processing, 2008.* 15_{th} *IEEE International Conference on* , 3112–3115 (October 2008).

[8] Hsu, Y.-F. and Chang, S.-F., "Detecting image splicing using geometry invariants and camera characteristics consistency," in [*Multimedia and Expo, 2006 IEEE International Conference on*], 549–552 (July 2006).

[9] Hsu, Y.-F. and Chang, S.-F., "Image splicing detection using camera response function consistency and automatic segmentation," in [*Proc. of International Conference on Image Processing*], 28–31 (2007).

[10] Chen, W., Shi, Y., and Su, W., "Image splicing detection using 2-d phase congruency and statistical moments of characteristic function," in [*Proc. of SPIE, Security, Steganography and Watermarking of Multimedia Contents IX*], (2007).

[11] Dong, J., Wang, W., Tan, T., and Shi, Y., "Run-length and edge statistics based approach for image splicing detection," in [*Digital Watermarking*], Kim, H.-J., Katzenbeisser, S., and Ho, A., eds., *Lecture Notes in Computer Science* **5450**, 76–87, Springer Berlin / Heidelberg (2009).

[12] Hsiao, D. and Pei, S., "Detecting digital tampering by blur estimation," in [*Proc. of 1st International Workshop on Systematic Approaches to Digital Forensic Engineering*], 264–278 (2005).

[13] Zhou, L. and Wang, D., "Blur detection of digital forgery using mathematical morphology," in [*Proc. of 1st KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*], 990–998 (2007).

[14] Goljan, M., Fridrich, J., and Filler, T., "Managing a large database of camera fingerprints," in [*Proceedings of SPIE: Media Forensics and Security II*], **7541** (January 2010).

[15] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transaction on* **13**, 600–612 (April 2004).

[16] Egiazarian, K., Astola, J., Ponomarenko, N., Lukin, V., Battisti, F., and Carli, M. in [*Two new full-reference quality metrics based on HVS*], *Proc. Second Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-06* (January 2006).

[17] Webster, A. A., Jones, C. T., Pinson, M. H., Voran, S. D., and Wolf, S., "Objective video quality assessment system based on human perception," in [*Proc. SPIE Vol. 1913, p. 15-26, Human Vision, Visual Processing, and Digital Display IV, Jan P. Allebach; Bernice E. Rogowitz; Eds.*], Allebach, J. P. and Rogowitz, B. E., eds., *SPIE* **1913**, 15–26 (September 1993).

[18] Bretillon, P., Baina, J., Jourlin, M., and Goudezeune, G., "Method for image quality monitoring on digital television networks," in [*Proc. SPIE Vol. 3845, p. 298-306, Multimedia Systems and Applications II, Andrew G. Tescher; Bhaskaran Vasudev; V. Michael Bove; Barbara Derryberry; Eds.*], Tescher, A. G., Vasudev, B., Bove, V. M., and Derryberry, B., eds., *SPIE* **3845**, 298–306 (November 1999).

[19] Wu, H. and Yuen, M., "A generalized block edge impairment metric for video coding," *Signal Processing Letters* **4**, 317–320 (November 1997).

[20] Wang, Z., Bovik, A., and Evans, B., "Blind measurement of blocking artifacts in images.," (2000).

[21] Caviedes, J. and Jung, J., "No-reference metric for a video quality control loop," in [*Proc. 5th World Multiconference on Systemics, Cybernetics, and Informatics*], 290–5 (2001).

[22] Wang, Z., Sheikh, H. R., and Bovik, A. C., "No-reference perceptual quality assessment of JPEGcompressed images," in [*Proc. IEEE International Conference on Image Processing*], (2002).