EXPLOITING PERCEPTUAL QUALITY ISSUES IN COUNTERING SIFT-BASED FORENSIC METHODS

I. Amerini[⋆] *F. Battisti*[†] *R. Caldelli*[⋆] *M. Carli*[†] *A. Costanzo*[*]

[⋆] Media Integration and Communication Center (MICC), Università degli Studi di Firenze, Firenze, Italy
[†] COMLAB Telecomm. Lab, Engineering Department, Università degli Studi Roma TRE, Roma, Italy
[*] Department of Information Engineering, University of Siena, Siena, Italy

ABSTRACT

Scale Invariant Feature Transform (SIFT) has been widely employed in several image application domains, including Image Forensics (e.g. detection of copy-move forgery or near duplicates). Recently, a number of methods allowing to remove SIFT keypoints from an original image have been devised studying the problem of SIFT security against malicious procedures. Such techniques are quite effective in producing an attacked image with very few (or no) keypoints, but at the expense of an image distortion. Final perceptual quality has been taken in account very roughly so far. In this paper, effectiveness of the attacking methods is evaluated also from the side of perceptual image quality; a new version of a SIFT keypoint removal method, based on a perceptual metric, is presented and an extended series of perceptive experiments is reported.

Index Terms— SIFT keypoint removal, counter forensics, image quality metrics, perceptual experiments.

1. INTRODUCTION

Recently an increasing number of forensic researchers explored the topic of counter-forensics, that is the study of methods to fool the forensic techniques by concealing the traces of manipulations [1]. Among the most common ways of manipulating the semantic content of a picture there is the copy-move forgery, whereby a portion of the image is copied once or more times elsewhere into the same image. Literature offers several examples of detectors for such manipulation [2]. Among them, recently [3, 4] those based on Scale Invariant Feature Transform (SIFT) [5] were proposed. The capability of SIFT to discover correspondences between similar visual content, in fact, allows the forensic analysis to detect very accurate and realistic copy-move forgeries. Furthermore, since SIFT is a powerful instrument to recognize and retrieve objects, an analysis on SIFT security becomes very important in order to assess if an attacker is able or not to succeed in deluding the image recognition process. All the studies on countering SIFT-based methods have demonstrated that devising procedures to attack SIFT is not a trivial task. SIFT features are not only robust against several non-malicious processing but also against tampering attempts. Most attacks, in fact, often alter the content in such a way that new valid keypoints are created and, more importantly, pay a high cost in terms of visual quality degradation. During the last years, research on counter-forensics has been mainly focused on the development of counter-forensic techniques that should be able to infer the related forensic methods. This spreading is accompanied by an increasing need for assessing the perceived quality of the resulting images. Interactions between perceived quality and security are more and more of interest [6]. Till now, only a rough analysis based on PSNR and SSIM metrics, was taken in account. Here to deeply understand the impact of the attacking methods on human perception a set of subjective tests has been performed. In this paper we present an analysis of SIFT keypoint removal from the perceptual quality point of view with the aim to improve the performance of the existing approaches. More specifically, we study the quality degradations of the attacked images produced by the algorithm in [7] and then we propose a Perceptual Keypoint Removal method based on PSNR-HVS-M metric [8] proving the improvement on the visual quality of the attacked image. For this reason many quality measures were evaluated and successively the PSNR-HVS-M was chosen to improve the attack in [7]. In the experimental results objective and subjective tests are made to compare the proposed method with others SIFT removal methods.

The rest of the paper is organized as follows. In Section 2 a brief overview of the counter-forensic methods is performed. In Section 3 the key elements of the proposed method are presented. In Section 4 the experimental results validating the system are reported and, finally, in Section 5 concluding remarks and future works are presented.

2. RELATED WORKS

The first attempt to test the security of SIFT has been made in [9]. In this early work, the authors were able to compromise an authentication system based on SIFT and image hashing by deleting keypoints. In 2010, Do *et al.* applied the technique

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

of [9] to assess the potential threat on a SIFT-based Content Based Image Retrieval (CBIR) scenario and demonstrated the robustness of their CBIR system to Hsu *et al.*'s attack. Following this analysis, Do et al. focused on the Content Based Image Retrieval (CBIR) scenario, devising new attacks to spatial locations [10] and to dominant orientations of keypoints and by showing that, in practical applications, concatenating multiple attacks may improve the final outcome. The topic of SIFT keypoints manipulation has been then investigated in image forensic and counter-forensic [1] scenarios in [11], where SIFT keypoints were removed by means of local warping attacks derived from image watermarking in order to impair SIFT-based detection of copy-move forgeries. The ideas of [11] were further developed in [7], where a new keypoint removal attack based on the classification of keypoints (see Section 3) was introduced. The work presented here has focused on the redefinition of the keypoint removal algorithm proposed in [7] and it studies the counter-forensic scenario on a perceptual quality metric point of view, which was not extensively evaluated so far.

3. KEYPOINT REMOVAL AND PERCEPTUAL ISSUES

In this Section we review the keypoint removal attack proposed in [7], called Classification-based Attack (in short, CLBA) and then a variation, based on a perceptual metric, is proposed. Such algorithm is based on the concept of keypoint classification preceding the attack itself; identifying classes of keypoints (unimodal, bimodal, multimodal) with different properties, in fact, allows to choose the attack that fits the most to such properties. Only the keypoints belonging to the first scale are considered. CLBA iterates the tasks of keypoint classification and tailored removal: given an input image I, for any iteration $k > 1$ the keypoints that were not removed are attacked again. The iterative procedure allows not only to remove more robust keypoints by progressively increasing the strength of the attack but also to deal with the not otherwise controllable side effect of removal [9, 11], that is the introduction of new keypoints that are very similar to those that have been deleted or the simple translation of old keypoints. CLBA halts when a certain condition is met, such as the maximum number of iterations max_iter or a minimum percentage of removed keypoints is reached. During the first half of the iterations, all the classes are attacked by means of Gaussian smoothing, which reduces the population of weaker keypoints without noticeable consequences on image quality. During the second half of the iterations, the surviving, more robust keypoints are deleted by means of Collage attack if they are unimodal or multimodal, and by means of Removal With Minimum Distortion (RMD) [10] if they are bimodal. The output of the attack is an image $J = \text{CLBA}(I)$ whose population of keypoints has been reduced up to (ideally) 0. In a nutshell, each attack works as follows. The Smoothing Attack is a simple light Gaussian filtering, whose usefulness in removing SIFT keypoints has been first observed in [10]. The Collage Attack has been employed with success in removing SIFT keypoints in [9]. In general, it consists on the substitution of an original patch with a new patch that is visually similar but should not contain any keypoints. In [7] such patch is drawn from a previously collected database of patches without keypoints and its histogram is at minimum distance from that of the original patch. The idea behind the RMD attack [10] is to calculate a patch ϵ that added to the neighborhood of a keypoint allows its removal.

To reduce the perceptibility of the attack and conceal the artifacts along the borders, CLBA blends the manipulated and original neighborhoods by means of the linear combination:

$$P_{new} = \alpha \cdot P_{\text{orig}} + (1 - \alpha) \cdot P_{\text{forged}}, \tag{1}$$

where: P_{orig} and P_{forged} are the original and the manipulated 8×8 patches respectively; α is a 8×8 weighting window whose elements are set to 1 along the patch borders and progressively decrease to 0 near the center.

3.1. Perceptual keypoint removal

The goal of our work was to modify the algorithm in [7], in order to improve the visual quality of the final image J. The first idea is to set a new halt condition to the algorithm: the maximum image quality degradation in addition to the maximum number of iterations and the percentage of removed keypoints. We decided to keep the classification step and the iterative procedure as in [7], then we calculated a quality metric score between the current patch P_{orig} and the new attacked patch P_{new} and if this value was above a certain threshold the attack was kept otherwise was refused. The method, in this fashion, was able to improve the quality of the final image J but there was a significant reduction of removed keypoints. So, for such a reason, we decided to devise a different strategy by maintaining the approach to perform classification-based attacks but classification is now made a-posteriori on the basis of a quality metric within the iterative procedure. At each iteration we compute the keypoints and then we attack each of them with Smoothing, Collage and RMD attacks ($att_i, i = 1, 2, 3$); the attack that produces a patch (in the neighborhood of the keypoint) with the best perceptual quality is selected. As evidenced in Equation 2, we calculated the quality metric $q(\cdot)$ between P_{orig} and P_{new}, for each of the three attacks, and then we select the attack that produce the maximum value

$$\arg \max_{i=1,2,3} (q(P_{\text{orig}}, P_{new,att_i})) \tag{2}$$

Finally, we substitute in the image the patch corresponding to the selected attack. The attack terminates after a certain number of iterations ($max\_iter = 40$) or when the desired percentage of deleted keypoints is reached (ideally 100%).

This strategy permits to improve the visual quality of the image avoiding annoying artifacts and achieving a satisfactory removal rate as highlighted in the experiments in Section 4.1.

4. EXPERIMENTAL RESULTS

In this Section experimental results are presented; in particular, in subsection 4.1 the criterion for the choice of the quality metric $q(\cdot)$ is explained and then a comparison between the proposed method and that one in [7], both in terms of keypoint removal and in terms of final perceptual quality, is presented. On the contrary, in subsection 4.2 the results obtained through a campaign of various perceptive tests are reported.

4.1. Objective quality assessment

To evaluate the proposed method we have gathered 20 images from UCID database [1]with size 512×384 pixel. We ran the two iterative methods described in Section 3 ($max\_iter = 40$), first the CLBA method [7] and then the proposed one. First of all, we evaluated the trend over the 40 attack iterations of 13 metrics (both perceptual and not) on the CLBA method [7], in order to choose the best measure to be used in the proposed approach (a list of the considered metrics is given in Table 1). At every iteration, the score of each metric is computed between the attacked (at that stage) image and the original one. The PSNR-HVS-M perceptual metric [8] has been selected because it is resulted to be as more sensible, according to the applied distortions, with respect to the others (see Figure 1, second row from the top), showing the highest variance (around 8) over all the iterations. In detail, the selected perceptual metric takes into account the contrast sensitivity function (CSF) and the model of visual between-coefficient contrast masking of DCT basis functions based on a human visual system (HVS). Furthermore it is proven, in [8], that the PSNR-HVS-M outperforms all the others metrics of the state of the art and demonstrates an appropriate correspondence to human perception.

After having chosen the PSNR-HVS-M metric for the proposed method, we have compared it with the technique in [7], both in terms of final image quality at last iteration between the original image and its attacked version and in terms of the final percentage of eliminated keypoints. Results have been averaged over all the images of dataset and we obtained the values written in Table 2. It can be observed that both methods achieved a high average removal rate though the CLBA attack, as expected, is able to delete a superior number of keypoints (+14.61%). Anyway, this is done at the expense of a strong impact on visual quality with respect to the proposed method. In particular, the perceptual method performs well both in terms of PSNR-HVS-M (+5.13 dB) and of PSNR (+2.42 dB); SSIM is similar and not so significant.

[1]UCID - An Uncompressed Colour Image Database

Table 1. Evaluated quality metrics.

Quality metric	Acronym
mean-squared error	MSE
peak signal-to-noise ratio	PSNR
structural similarity index	SSIM
multiscale SSIM index	MSSIM
visual information fidelity	VIF
pixel-based VIF	VIFP
universal quality index	UQI
image fidelity criterion	IFC
noise quality measure	NQM
weighted signal-to-noise ratio	WSNR
signal-to-noise ratio	SNR
PSNR human visual system (DCT)	**PSNR-HVS-M**
PSNR human visual system	PSNR-HVS

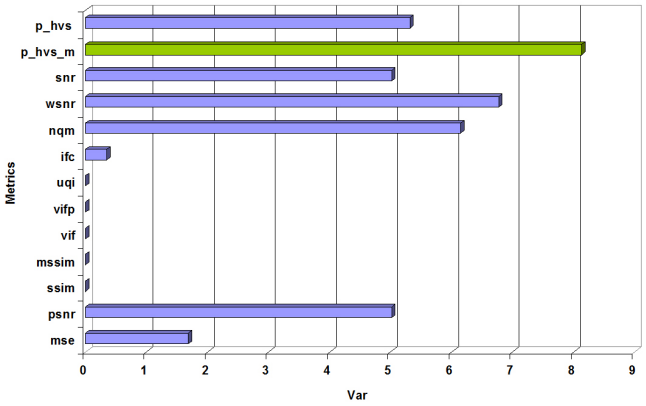

Fig. 1. Variance of the evaluated quality metrics over 40 iterations.

Table 2. Performance comparison between keypoint removal attacks.

	CLBA [7]	Proposed
Kpt removal rate	**97.49** %	82.88 %
PSNR-HVS-M	42.21 dB	**47.34 dB**
PSNR	45.12 dB	**47.54 dB**
SSIM	0.996	**0.997**

4.2. Subjective quality assessment

In order to assess the quality improvement due to the use of a perceptual model, we performed four campaigns of subjective tests. In these experiments, the subjective scores have been collected by means of in presence and crowdsourcing-based [12] tests. Crowdsourcing allows to reduce the costs and time needed for performing the experiments but the crowd is a large mass of different anonymous individuals and, for this reason, not always reliable. To cope with this problems we decided to perform some of the experiments in presence and

some through crowdsourcing. The first experiment (*Test1*) was devoted to verify the impact of the proposed method on the perceived image quality. 250 subjects participated to a crowdsourcing-based campaign. After the screening process performed for removing outliers and incomplete results, a set of 213 subjects were considered. Each subject evaluated 60 different images: 10 original images and 50 modified ones produced by the proposed perceptual method at different iterations (*iter* = 1, 2, 3, 5, 40). The test was performed by using an Absolute Category Rating with Hidden Reference approach (ACR-HR) [13]. These images were presented, in random order, one at a time for 6 seconds and were rated independently on a scale from 1 to 5 where 1 equals poor quality and 5 excellent quality. The size of the images is 512×384 pixels and the zoom is not allowed. The obtained results are reported in Table 3. It can be noticed that the average MOS (Mean Opinion Score) does not change significantly with the number of considered iterations. Moreover, the results show that the subjects were not able to discriminate between originals and attacked images thus proving that the perceptual method does not significantly affect the image quality.

Table 3. Result of the first subjective test (*Test1*).

Image	Average MOS
Original	3.502
Attacked iter 1	3.531
Attacked iter 2	3.553
Attacked iter 3	3.521
Attacked iter 5	3.488
Attacked iter 40	3.507

In order to further understand the impact of the algorithm on the perceived quality, a second experiment (*Test2*) has been performed. Twenty-five subjects took part at *Test2*. In this test, a Pair Comparison (PC) approach has been selected [13]. The stimuli were displayed on a Panasonic BT-3DL2550 screen (1920×1080 pixels). Each subject was asked to choose, for every couple of images displayed on the screen, the one that according to him/her was modified (i.e. attacked by the proposed method at $iter = 1, 2, 3, 5, 40$). For each couple, the original image and the modified one were randomly displayed on the left and right side of the screen. The results are depicted in Figure 2 where correct detections are represented with respect to the users. It is possible to notice that it is a complicate task for users (also in a controlled environment) to distinguish the original image from the attacked one demonstrating the effectiveness of the proposed method.

In the third test (*Test3*), we wanted to evaluate the proposed method ($iter = 40$) by comparing it with the CLBA method [7] in a controlled environment. 10 couple of images are displayed for each of 54 subjects. The subjects were asked to select which of the two images was preferable from a visual quality point of view. The proposed method was preferred in

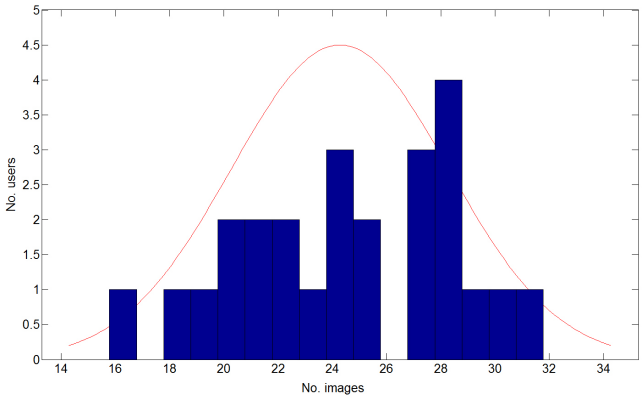

Fig. 2. Results of the second experiment (*Test2*).

the 65% of cases with respect to the CLBA.

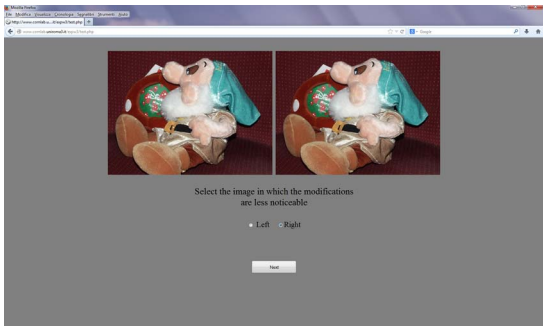

Fig. 3. Graphical user interface used in *Test4*.

Finally, a fourth experiment, *Test4*, has been performed, exploiting crowdsourcing, as shown in Figure 3. A number of 50 subjects participated to the tests. They were asked to select the image in which the modifications are less noticeable between an image modified with the proposed method (as before) and with the RMD [10], included in an iterative procedure. The 90% of the subjects considered the proposed method better than RMD.

5. CONCLUSIONS

In this paper we presented an evaluation of a counter-forensics scheme to fool a SIFT-based technique from the side of perceptual image quality. Furthermore, a new version of a SIFT keypoint removal method, based on a perceptual metric, is presented and a series of perceptive experiments is reported. We demonstrated that the proposed method obtains the lowest possible impact on visual quality with respect to the methods presented so far still achieving to remove a relevant number of keypoints. In the future could be studied the perceptual quality issue in the case of the injection of fake keypoints into an attacked image increasing the number of pictures in the dataset.

6. REFERENCES

[1] R. Böhme and M. Kirchner, "Counter-forensics: attacking image forensics," in *Digital Image Forensics (Chapter 10)*, Husrev Taha Sencar and Nasir Memon, Eds. Springer, New York, 2012.

[2] M.C. Stamm, Min W., and K.J.R. Liu, "Information forensics: An overview of the first decade," *Access, IEEE*, vol. 1, pp. 167–200, 2013.

[3] X. Pan and S. Lyu, "Region duplication detection using image feature matching," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 857–867, 2010.

[4] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A SIFT-based forensic method for copy move attack detection and transformation recovery," *Information Forensics and Security, IEEE Trans. on*, vol. 6, no. 3, pp. 1099 –1110, sept. 2011.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int.'l Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] F. Battisti, M. Carli, and A. Neri, "Image forgery detection by means of no-reference quality metrics," in *Proc. of SPIE - Media Watermarking, Security, and Forensics*, 2012.

[7] I. Amerini, M. Barni, R. Caldelli, and A. Costanzo, "Counter-forensics of SIFT-based copy-move detection by means of keypoint classification," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–17, 2013.

[8] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2007.

[9] C-Y Hsu, C-S. Lu, and S-C Pei, "Secure and robust SIFT," in *Proc. of the 17th ACM international conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 637–640, ACM.

[10] T-T. Do, E. Kijak, T. Furon, and L. Amsaleg, "Deluding image recognition in SIFT-based CBIR systems," in *Proc. of the 2nd ACM workshop on Multimedia in forensics, security and intelligence*, 2010, MiFor '10, pp. 7–12.

[11] R. Caldelli, I. Amerini, L. Ballan, G. Serra, M. Barni, and A. Costanzo, "On the effectiveness of local warping against SIFT-based copy-move detection," in *Proc. of Int'l Symposium on Communications, Control and Signal Processing (ISCCSP)*, Roma, Italy, May 2012.

[12] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd. A framework for crowd-based quality evaluation," in *Proc. of the IEEE Picture Coding Symposium*, 2012, pp. 254–248.

[13] ITU, "Subjective video quality assessment methods for multimedia applications," in *Recommendation P.910*, 2008.