ON THE EFFECTIVENESS OF LOCAL WARPING AGAINST SIFT-BASED COPY-MOVE DETECTION

R. Caldelli, I. Amerini, L. Ballan, G. Serra

Media Integration and Communication Center
University of Florence
Viale Morgagni, 65, Florence, Italy

M. Barni, A. Costanzo

Department of Information Engineering
University of Siena
Via Roma, 56, Siena, Italy

ABSTRACT

One of the simpler and most used method to alter the content of a digital image is to copy-move a portion of it onto another area with the intent, usually, to hide something awkward. In image forensics scientific community, this kind of modification is generally detected by resorting at techniques based on SIFT features that provide a local description which is robust to global geometric transformations the image may undergo. On such a basis, this paper investigates the effectiveness of some methodologies which introduce a local warping onto the copy-pasted patches in order to reduce the detection capability of SIFT-based approaches. This analysis is particularly interesting in a real scenario of forensic security. Four diverse local warping techniques have been taken into account and experimental results with respect to final perceptual quality of the forged image are presented.

Index Terms— Local warping, SIFT, copy-move, forensic security.

1. INTRODUCTION

Nowadays, local visual features are widely used for image retrieval and object recognition, due to their robustness to several geometrical transformations (such as rotation and scaling), occlusions and clutter. More recently, attempts have been made to apply these kinds of features also in the image forensics domain, in particular to understand if a digital image is authentic or has been counterfeited. Specifically, such local features are used to detect copy-move forgeries. In fact, this is one of the most common image manipulations in which an area of an image is copied and then pasted onto another zone to duplicate or conceal something in the scene. Most of the known algorithms which try to detect such infringement are based on SIFT (Scale Invariant Features Transform) descriptors which allow to match image keypoints presenting similar features [1, 2].

Recently, the discussion on the reliability of image forensic algorithms has gained importance by starting also to consider what an attacker could try to do to invalidate a specific forensic technique. To study the possible attacks to deceive the image forensic methods and find some countermeasures are two key issues in the forensic security field [3, 4].

Furthermore, an analysis on SIFT security has not been performed yet. In [5] the impact of very specific attacks against the SIFT descriptors is analyzed on a real CBIR (Content-Based Image Retrieval) system in order to succeed in deluding the image recognition process. However the security issue of SIFT is relatively unexplored especially in the forensic scenario of copy-move detection. In particular, being well-known that SIFT features perform efficiently against global geometric transformation, it is worthy to be analyzed their behavior in front of local distortions. On such a basis, this paper investigates the potential weaknesses of the SIFT-based forensic method for copy-move attack detection versus different local warping methods that an attacker could implement in performing his illegal action.

The paper is structured as follows: Section 2 reviews the SIFT technique and the copy-move detection method and Section 3 presents the analyzed local warping attacks. In Section 4 experimental results are presented and Section 5 concludes the paper.

2. SIFT AND COPY-MOVE DETECTION

In this Section a brief review of the SIFT technique and of the approach for detecting copy-move forgeries is drawn.

Given an image, SIFT features [6] are detected at different scales by using a scale space representation implemented as an image pyramid. The pyramid levels are obtained by Gaussian smoothing and sub-sampling of the image resolution while interest points are selected as local extrema (min/max) in the scale-space. These points (usually called *keypoints*) are extracted by applying a computable approximation of the Laplacian of Gaussian (LoG) called Difference of Gaussians (DoG). In particular, the SIFT algorithm approximates LoG by iteratively computing the difference between two nearby scales in the scale-space. Once these keypoints are detected, SIFT descriptors are computed at their locations in both image plane and scale-space. Each descriptor consists in a histogram of 128 elements, obtained from a 16x16 pixels area

around the corresponding keypoint. The contribution of each pixel is obtained by calculating image gradient magnitude and direction in scale-space and the histogram is computed as the local statistics of gradient directions (8 bins) in 4x4 sub-patches of the 16x16 area.

The procedure in which interest points are localized ends with a list of N keypoints each of which is completely described by the following information: $\mathbf{x}_i = \{x, y, \sigma, o, \mathbf{f}\}$, where (x, y) are the coordinates in the image plane, σ is the scale of the keypoint (related to the level of the image-pyramid used to compute the descriptor), o is the dominant orientation (used to achieve rotation invariance) and $\mathbf{f}$ is the final SIFT descriptor.

After SIFT features are extracted the copy-move forgery detection is performed in the SIFT space among the $\mathbf{f}_i$ vectors of each keypoint to identify similar local patches in the test image. The best candidate match for each keypoint $\mathbf{x}_i$ is found by identifying its nearest neighbor from all the other $(n-1)$ keypoints of the image, which is the keypoint with the minimum Euclidean distance in the SIFT space. In order to decide if two keypoints match the ratio between the distance of the closest neighbor to that of the second-closest one is used, and then this ratio is compared with a threshold T (typically fixed to 0.6). For the sake of clarity, given a keypoint we define a similarity vector $\mathbf{D} = \{d_1, d_2, \ldots, d_{n-1}\}$ that represents the sorted euclidean distances with respect to the other descriptors. The keypoint is matched only if this constraint is satisfied:

$$d_1/d_2 < T \quad where \quad T \in (0, 1). \tag{1}$$

Finally, by iterating over keypoints in $\mathbf{X}$, we can obtain the set of matched points which, at this stage, already provides a draft idea of the authenticity of the image and of the presence of duplicated areas (see [1] for further details). Procedures of segmentation and clustering can successively be adopted to better individuate manipulated patches.

3. LOCAL WARPING ATTACKS

The idea, proposed in this work, is to counter-attack forensic techniques which detect copy-move forgeries by resorting to a SIFT-based approach. Specifically, the counter-action is exploited by using algorithms of local warping to reduce performances of such methods. The procedure is the following:

1. Create a copy-moved forged image (F)
2. Select source (S) and destination (D) patches
3. Apply a local warping method to both S and D
4. Paste back the warped patches onto image F by obtaining the final image F_{LW}

In our analysis, we would like to check how different local warping techniques behave in this application scenario; in particular, we have taken into account four well known methods which will be briefly described in the sequel (multiple copy-move are not considered at this stage).

3.1. StirMark

The StirMark software [7] is very well-known within digital watermarking scientific community as a local random bending attack to de-synchronize watermark extraction systems. It basically comprehends a sequence of three transformations, whose the second applies a displacement which is zero at the border of the image and maximum at the center (it depends upon the parameter b), while the third, the actual local distortion, applies a random displacement at each pixel location according to the parameter R. Perceptual quality of the stirmarked image was usually very high but watermark recovery was inhibited.

3.2. LPCD

The second method taken into consideration is the Local Permutation with Cancelation and Duplication (LPCD) [8]. Let Δ define a discrete set whose values are integer numbers in $I = [-\Delta, \Delta]$. LPCD modifies an image according to the following rule: if $B(i, j)$ is a generic pixel of the distorted image B, let $B(i, j) = A(i + \Delta_h(i, j); j + \Delta_v(i, j))$ where A is the original image and $\Delta_h(i, j)$ and $\Delta_v(i, j)$ are i.i.d. integer random variables uniformly distributed in I. Such modification does not introduce block artifacts because of the overlapping of the windows of the possible displacements for neighboring pixels. A multiresolution version of this model has been proposed to improve the obtainable perceptual results. Let $n \times m$ be the size of the image and L the resolution value: a low resolution displacement field of size $\frac{n}{2^L} \times \frac{n}{2^L}$ is first generated, then a full size displacement field is obtained by means of bilinear interpolation. The full resolution field is applied to the original image to produce the distorted image. Therefore, the parameters to be controlled are L and Δ.

3.3. C-LPCD

The LPCD model does not contain any constraint on the smoothness of the displacement field, so there is no guarantee that the set of applied distortions is perceptually invisible, even by considering very small values of Δ. So in the C-LPCD (Constrained LPCD) has been added the requirement that the sample order is preserved thus introducing memory in the system. In other words, the horizontal and vertical displacements of the pixel (i, j) are limited by the horizontal and vertical displacements of the pixels $(i - 1, j)$, $(i, j - 1)$ and $(i - 1, j - 1)$. The parameters to be controlled are the same of the LPCD model.

3.4. Markov Random Fields

The last local warping method we considered is based on Markov Random Fields (MRF) [9]. The objective is to generate a displacement field according to a defined Gibbs probability distribution and to a specific potential function. Firstly, the displacement field is initialized by assigning to each pixel (i, j) in the image a displacement vector $\mathbf{f}(i, j)$ generated randomly (and independently from the other pixels) whose magnitude is determined by relying on perceptual considerations. Such initial random field is treated as a noisy version of an underlying displacement field obeying the MF model. The MF field is then obtained by applying an iterative smoothing algorithm to the randomly generated one. The technique randomly visits all the points of the displacement field and updates their values by trying to minimize the defined potential function. After that each pixel has been visited and the corresponding displacement updated, a new iteration starts. The algorithm ends when no new modification is introduced for a whole iteration. As for LPCD and C-LPCD, better visual results can be achieved by means of multiresolution MRF. The main parameters controlling the method are then the resolution L at which the displacement field is created and the standard deviation σ of the employed potential function.

4. EXPERIMENTAL ANALYSIS

In this Section the parameters of the local warping algorithms of Section 3 are introduced, then the effectiveness against a copy-move detection method based on SIFT is investigated.

4.1. Experimental results

To assess the impact of local warping attacks on SIFT features we focused on a dataset of 10 JPEG color images whose size ranged from 700×500 to 1000×800 pixels. Amongst them we have also included the same image used by Pan et al. in [2], which is shown in Figure 1(a). Each image of the dataset has undergone a realistic copy-move forgery under the hypothesis that only two patches are present in the resulting fake, as shown in Figure 1(b). The size of the copy-moved patches ranged approximately from 100×100 to 250×250 pixels. The manipulation has been carried on by means of Adobe Photoshop®. We did not resort to any post-processing effects because they may affect the matches between keypoints of the two patches, thus risking to alter the final results of our study. The forensic technique we used to detect the copy-move forgery is the one proposed by Amerini et al. [1], integrated with the Vedaldi's implementation of SIFT algorithm [10] and with the parameters set as suggested by Lowe [6]. By applying these methods to the dataset of forged images the detection method found a matches between copy-moved patches that ranged from 21 to 96.

Although the performance of the warping methods may be influenced by several parameters, in order to keep the analysis under control we have focused only on two amongst the most important for each of them: (L,Δ) for LPCD and C-LPCD; (L,σ) for MRF; and (b,R) for StirMark. By assigning different values to these pairs we can effectively control the strength of the attack. Two examples of the results one can obtain are shown in Figure 2: the image on the left has been produced by means of LPCD with parameters (L,Δ)=(1,4), while the one on the right by means of StirMark (v3.1) with parameters (b,R)=(4,0.9).

(a) Original image (b) Copy-move forgery

Fig. 1. One of the images used for the experiments: yellow shapes indicate the copy-move forgery. The detection technique of [1] found 52 matches between copy-moved patches.

Fig. 2. Examples of attacked images. Left: LPCD with (L,Δ)=(1,4). Right: StirMark 3.1 with (b,R)=(4,0.9).

More specifically, in our experiments L ranged in the integer interval [1,6] (step=1), Δ in [3,7] (step=1), σ in [1,9] (step=2), b in [6,12] (step=2) and R in [0.1,1.5] (step=0.2). These values have been assigned accordingly with those suggested by the theory underlying the algorithms and for each forged image produced 30 attacked versions for LPCD, C-LPCD and MRF and 32 attacked versions for StirMark.

The evaluation of the local warping impact has been performed by means of the following two measures: (i) the percent of matches between copy-moved patches that were eliminated by the attack; (ii) the average visual quality over the two warped patches compared to their counterparts prior to the attack. We have computed such quality by means of the Gabor metric proposed by D'Angelo et al. [11], which was designed to overcome the limitations of classical indices such as PSNR or SSIM in rating the perceptual effect of local geometric attacks. Given an input image and the displacement field of the warping attack it has undergone, the Gabor metric provides a score in the scale $[1, 5] \in \mathbb{R}$, where values range from "bad" quality (near 1) to "excellent" quality (near 5). For example, the warping on images of Figure 2 removed respectively the 100% and the 89% of total matches between

patches with a Gabor score of 1.03 and 1.91.

For each algorithm these two measures have been related to each other as follows. Given a forged image, we first sampled the percent of eliminated matches, thus obtaining the Gabor metric's score of all those attacked versions that achieved a specific percent of elimination. Then we computed the global score as the average of such Gabor metrics (missing values have been obtained by means of cubic interpolation). The final results shown in Figure 3 correspond to the average of this procedure over the 10 images composing our test dataset.

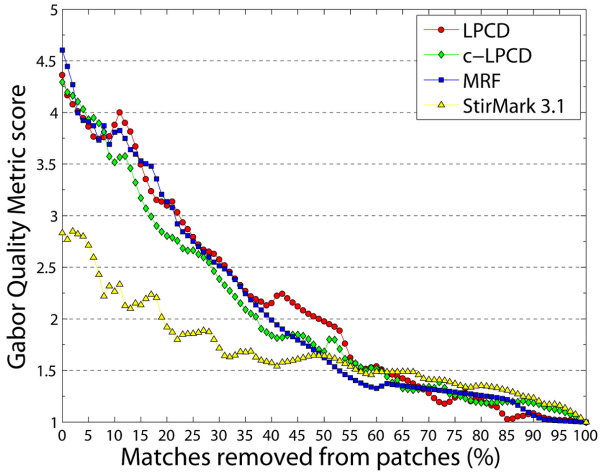

Fig. 3. Average Gabor quality metric depending on the percent of removed matches.

As expected, a trade-off exists between the amount of removed matches and the perceptual quality: the stronger the attack the higher the amount of matches that are removed. This result, however, is achieved at the cost of a loss of visual quality. If we focus on the left portion of the curves (0-50% removed matches) we can see that StirMark is clearly outperformed by the others methods, whose behavior is substantially comparable. On the right portion (50-100% removed matches), however, it is StirMark (along with MRF) that provides the best possible quality. Although the quality may appear low, it is important to point out that it has been calculated only on the attacked patches. If the local warping is not excessively strong, the overall quality of the image remains acceptable.

In Figure 4, an example of matched keypoints reduction is presented. It is possible to see that the number of SIFT matches between the source and destination patches is drastically diminished going from the initial number of 52 to just 8 (a loss of 88.5%) when a local warping attack (in this case LPCD with (L,Δ)=(3,3)) is applied before pasting. The global perceptual quality of the image is not affected at all, though a Gabor score of 1.43 is obtained over the two warped areas with respect to the un-warped case.

Fig. 4. Example of loss of SIFT matches: copy-move forgery (left) and LPCD with (L,Δ)=(3,3) (right). The initial 52 matches are reduced to 8 (88.5% are eliminated) with a Gabor score of 1.43.

In our analysis we mainly focused on images with landscape content. We noticed, in fact, that when the image content is characterized by many vertical or horizontal edges (e.g. buildings, walls) all the methods, although still effective in removing matches, do not provide good visual results. The reason behind this is that the warping strength becomes clearly visible along the regular edges, thus producing effects that are perceptually more disturbing. An example is provided in Figure 5: in this case, the copy-move forgery prior to the warping (left) is characterized by 96 matches between keypoints belonging to the patches. The C-LPCD attacked version (right) removed the 60% of matches with a Gabor score equal to 1.

Fig. 5. Example of local warping on an image with regular edges: copy-move forgery (left) and C-LPCD with (L,Δ)=(1,6) (right). The effects of distortions are clearly visible (60% matches removed with Gabor score 1).

5. CONCLUDING REMARKS

In this paper an analysis on the effectiveness of local warping attacks against SIFT-based copy-move detection is presented. To assess the impact of local attacks on SIFT features we compared four different local warping algorithms in terms of removed matches after the attack and visual quality of the attacked patches. We showed that an increment of the attack strength coupled with an augment in the number of removed matches. A complete keypoints removal is possible and is achieved, as expected, at the cost of a loss of visual quality. The results of the attack also depend on the content of the image: distortions, in fact, appear to be more disturbing on images with regular edges (e.g. buildings). Several aspects of the problem could be further investigated, such as: study of ad-hoc warping techniques that preserve regular edges; in-

clusion of more than two copy-moved patches; analysis on a larger number of copy-moved images.

6. ACKNOWLEDGMENT

This work was partially supported by the REWIND Project funded by the Future and Emerging Technologies (FET) programme within the 7FP of the European Commission, under FET-Open grant number: 268478.

7. REFERENCES

[1] 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 Transactions on*, vol. 6, no. 3, pp. 1099 –1110, sept. 2011.

[2] 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.

[3] M. Goljan, J. Fridrich, and Mo Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 1, pp. 227 –236, 2011.

[4] R. Caldelli, I. Amerini, and A. Novi, "An analysis on attacker actions in fingerprint-copy attack in source camera identification," in *IEEE Workshop on Information Forensics and Security*, 2011.

[5] Thanh-Toan Do, Ewa Kijak, Teddy Furon, and Laurent Amsaleg, "Deluding image recognition in sift-based cbir systems," in *Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence*, New York, NY, USA, 2010, MiFor '10, pp. 7–12, ACM.

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

[7] F. Petitcolas, R. Anderson, and M. Kuhn, "Attacks on copyright marking systems," in *Proc. of Information Hiding, Second International Workshop, IH98,*, Portland, Oregon, USA, 1998, pp. 219–239.

[8] A. D'Angelo, M. Barni, and G. Menegaz, "Perceptual quality evaluation of geometric distortions in images," in *Proc. of of SPIE Human Vision and Electronic Imaging XII*, 2007, vol. 6492.

[9] S. Li, *Markov random field modeling in computer vision*, Springer-Verlag, London, UK, 1995.

[10] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.

[11] A. D'Angelo and M. Barni, "A structural method for quality evaluation of desynchronization attacks in image watermarking," in *IEEE 10th Workshop on Multimedia Signal Processing*, Oct. 2008.