Understanding the Security and Robustness of SIFT

Thanh-Toan Do, Ewa Kijak, Teddy Furon, Laurent Amsaleg

Understanding the Security and Robustness of SIFT

Thanh-Toan Do
Université de Rennes 1, IRISA
thanh-toan.do@irisa.fr

Ewa Kijak
Université de Rennes 1, IRISA
ewa.kijak@irisa.fr

Teddy Furon
INRIA Rennes, IRISA
teddy.furon@irisa.fr

Laurent Amsaleg
CNRS, IRISA
IRISA - Campus de Beaulieu,
35042 Rennes cedex
laurent.amsaleg@irisa.fr

ABSTRACT

Many content-based retrieval systems (CBIRS) describe images using the SIFT local features because of their very robust recognition capabilities. While SIFT features proved to cope with a wide spectrum of general purpose image distortions, its security has not fully been assessed yet. In one of their scenario, Hsu *et al.* in [2] show that very specific anti-SIFT attacks can jeopardize the keypoint detection. These attacks can delude systems using SIFT targeting application such as image authentication and (pirated) copy detection.

Having some expertise in CBIRS, we were extremely concerned by their analysis. This paper presents our own investigations on the impact of these anti-SIFT attacks on a real CBIRS indexing a large collection of images. The attacks are indeed not able to break the system. A detailed analysis explains this assessment.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Security

1. INTRODUCTION

Content-based image retrieval systems (CBIRS) are now quite mature. They typically use advanced multidimensional indexing technique and powerful low-level visual descriptors, making them both efficient and effective at returning the images from a database that are similar to query images. An abundant literature describes such systems and evaluates their ability to match images despite severe distortions [3]. Such good recognition capabilities are essentially due to distinctive local features computed over images, the most popular example being the SIFT descriptors [5]. CBIRS are used in many applications, including security oriented scenarios like copyright enforcement and image forensics.

So far, it is mostly the *robustness* of CBIRS that has been evaluated. In contrast, very few works investigate their security level. *Security* of CBIRS is challenged when pirates mount attacks after having accumulated a deep knowledge about a particular system, focusing on very specific parts where flaws have been identified. This security perspective recently gained interest in the multimedia community.

In 2009, a paper by Hsu *et al.* [2] discusses extremely specific anti-SIFT attacks potentially making it hard for a CBIRS to subsequently match the attacked images with the ones from the database. In a copy detection scenario, this would allow illegal material to remain undetected. This is an extremely serious threat, as so many real life systems use such visual features. Hsu *et al.* use this claim to motivate an encryption mechanism securing the SIFT description. We immediately decided to rerun their experiments against our own system, in realistic settings, to assess the seriousness of their conclusions.

Surprisingly, after carefully implementing their anti-SIFT attacks and running similar experiments, we were able to draw very different conclusions. While [2] suggests a SIFT-based system can be broken, we could not break our system, that is, attacked images were still matched with their original version. We then decided to undertake a deep investigation of the phenomenons at stake.

This paper makes the following contribution: it provides an in depth analysis of what happens when performing the "Keypoint Removal" attack, which is central to one scenario in [2]. In particular, this paper shows removing keypoints triggers the creation of new and undesired keypoints that are easy to match.

This paper is structured as follows. Section 2 details the material from [2] that is needed to understand the remainder of this paper. Section 3 describes our set up and the real life experiments we did, which turned out to contradict [2]. Section 4 describes in details what happens when removing keypoints as in [2]. That section also provides quite extensive details on the new and undesired keypoints created as a side effect of removals. Section 5 concludes the paper.

2. ROBUST AND SECURE SIFT

The main contribution of [2] is a secure implementation of the SIFT description. To motivate this, its authors first

MM'10, October 25–29, 2010, Firenze, Italy.

exhibit specific attacks modifying the image only around the detected keypoints to conceal them.

A keypoint (x, y, σ) is detected when it yields a local extremum over its neighborhood of the DoG (Difference of Gaussians) function $D(x, y, \sigma)$, which is the convolution of the image I by the difference $G_{d,\sigma}$ of two Gaussian smoothing kernels at scales σ and $k\sigma$: $D(x, y, \sigma) = G_{d,\sigma} \star I(x, y)$. We denote by $D(x_1, y_1, \sigma)$ a local extremum value, and by $D(x_2, y_2, \sigma)$ the second extremum in its spatial neighborhood. The following difference d is thus positive if the extremum is a maximum (resp. negative for a minimum):

$$\begin{aligned} d &= D(x_1, y_1, \sigma) - D(x_2, y_2, \sigma) \\ &= \sum_{(u,v)\in\mathcal{R}_1} I(u, v) G_{d,\sigma}(u - x_1, v - y_1) \\ &- \sum_{(u,v)\in\mathcal{R}_2} I(u, v) G_{d,\sigma}(u - x_2, v - y_2), \end{aligned}$$

where $\mathcal{R}_i$ is the support of the convolution kernel $G_{d,\sigma}$ translated by (x_i, y_i), $i \in \{1, 2\}$.

The idea of [2] is to locally add a patch ϵ to create the attacked image $I_m = I + \epsilon$, such that d is now null when calculated on I_m. There is no longer a unique local extremum at this point and therefore the keypoint has been successfully removed. The authors propose the following patch:

$$\epsilon(x, y) = \begin{cases} -d/|\mathcal{D}_1| G_{d,\sigma}(x - x_1, y - y_1) & \text{if } (x, y) \in \mathcal{D}_1 \\ d/|\mathcal{D}_2| G_{d,\sigma}(x - x_2, y - y_2) & \text{if } (x, y) \in \mathcal{D}_2 \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{D}_i = \mathcal{R}_i - \mathcal{R}_1 \cap \mathcal{R}_2$, $i \in \{1, 2\}$, $|\mathcal{D}_i|$ is its size in pixel. A ratio $\alpha > 0$ controls the keypoint removal rate: the patch is applied if $|\epsilon(x, y)| \leq \alpha I(x, y)$, $\forall (x, y) \in \mathcal{D}_1 \cup \mathcal{D}_2$.

Some experimental results measure the efficiency of the attack by evaluating its distortion by the PSNR between I and I_m when a certain percentage of the keypoints have been removed (from 90 to 10%) over three images (Lena, Baboon, and Bridge). Their secure implementation processes the image with a secret keyed transformation prior to the keypoint detection. This prevents the above attack since the pirate ignoring this secret can no longer locate the keypoints. A final experiment assesses that the robustness of the keypoint detector is not affected by the secret keyed transform. To this end, a database of 1,940 modified images (using the Stirmark benchmark) is built. When an original image is queried, their system almost always retrieved all the corresponding modified images. Another experiment shows how to defeat an image authentication scheme based on robust hash by inserting tampered areas having similar SIFT keypoints. We are not analyzing this last scenario since our paper only focuses on threats upon CBIR systems.

3. REAL TESTS AND CONTRADICTIONS

This section describes our experiments using a real CBIRS when trying to reproduce the results of [2]. We therefore carefully implemented their keypoint removal strategy, referred to as KPR-based attacks.

3.1 Algorithms for Description and Indexing

We computed the local SIFT descriptors using the open-source SIFT-VLFeat code by Vedaldi [6]. We did several experiments to get descriptors that are as close as possible to the original SIFT computed using Lowe's binary, both in terms of number of descriptors and of spatial location in images. In our case, the best configuration is when peak-thresh=4 and edge-thresh=12.

All the descriptors of the image collection are then processed by the NV-Tree high-dimensional indexing scheme [4]. The NV-Tree runs approximate k-nearest neighbor queries and has been specifically designed to index large collections of local descriptors.

3.2 Dataset and Queries

Our image collection is composed of 100,000 random pictures downloaded from Flickr that are very diverse in contents. All images have been resized to 512 pixels on their longer edge. This collection yields 103,454,566 SIFT-VLFeat descriptors indexed by the NV-Tree. We then randomly picked 1,000 of these images and performed 7 KPR-attacks on them with $\alpha \in \{0.2, 0.4, 0.8, 1.5, 3, 6, 12\}$.[1] For comparison, we also applied 49 standard Stirmark attacks (rotations, crops, filters, scalings, affine transforms, ...). Overall, there are 56,000 queries distributed in 56 families. This experimental protocol clearly targets a copy detection scenario.

3.3 Copy Detection Experiments

For all queries we kept track of the scores of the 100 best matching images. Fig. 1 illustrates the outcome of this experiment by selecting the 7 KPR-attacks as well as 3 Stirmark attacks. The score of the original image decreases as fewer keypoints remain after the attacks. The score of the best non matching image does not vary as it corresponds to (a stable number of) random matches. It clearly shows that the system works for all attacks.[2] Note also that the original images are always found when querying with any of the KPR-based attacked images. In other words, it is not possible to conceal images from the system when performing the anti-SIFT attacks of [2]. Indeed, this surprising result is backed-up by the facts revealed in the following analysis.

4. ANALYSIS

This section investigates why we were not able to confirm the results of [2] in a CBIRS context. We closely look at what happens when performing KPR-based attacks, and explain why attacked and original images still match. We use the Lena image as a vehicle for providing explanations; we observed and report similar phenomenons when using 1,000 random images, however.

4.1 Removal of Keypoints

Let us apply the patch ϵ of Sect. 2 at a particular position (x_1, y_1) of the Lena image. It was originally detected as a keypoint because of its local maximum at scale $\sigma = 1.37$. Fig. 2(a) shows the DoG local extrema originally detected, identified on the figure by a blue square, and the second extrema in its neighborhood (green circle). After attack (x_1, y_1) is no longer an extremum as the original first and second local extrema values are now quite equals (Fig. 2(b)).

Fig. 3(a) shows what this particular patch looks like in the neighborhood $\mathcal{R}_1 \cup \mathcal{R}_2$ of (x_1, y_1). It introduces strong

[1]These values are the ones best reproducing the keypoint removal rates used in [2].

[2]None of the Stirmark attack produced concealment.

Table 1: Number of deleted and new created keypoints after KPR-attacks for different values of keypoint removal ratio α on Lena and Baboon images, and average values computed over the 1,000 query images.

Image name	Total # of KP	α	% KP deleted	# KP deleted	# KP created	# KP after attack	% KP created in attacked image	PSNR
LENA	1218	0.2	20%	245	188	1161	16%	48.98
LENA	1218	12	90%	1093	756	881	86%	32.49
BABOON	3124	0.2	19%	590	418	2952	14%	43.24
BABOON	3124	12	92%	2865	1639	1898	86%	27.51
AVG ON 1000 IM	1034	0.2	14%	149	109	994	11%	50.81
AVG ON 1000 IM	1034	12	84%	871	605	768	79%	32.42

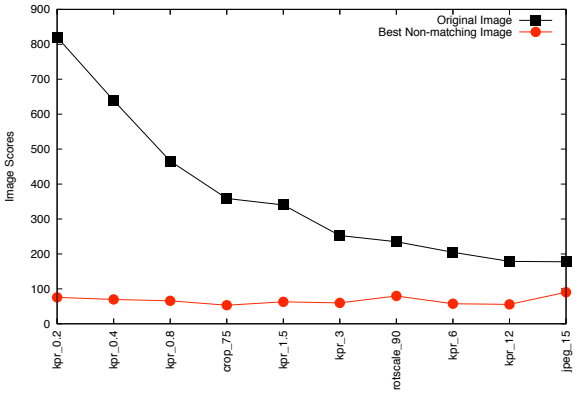


Figure 1: Image scores in realistic settings. X-axis: 10 selected attacks. Y-axis: for each family of attacks, the average scores over the 1,000 queries of the original images (expected to be matched) and of non matching images having the highest score.

Table 2: Average distance per octave between original and new keypoint location in Lena image.

	# KP deleted	# KP created	avg dist
Octave -1	812	483	3.0
Octave 0	197	190	2.6
Octave 1	580	550	3.7
Octave 2	210	220	3.2

visual distortion in the image (see Fig. 3(b)). This is not surprising since the patch is proportional to the inverse of the DoG kernel which vanishes at the edge of its support.

4.2 Removal Triggers Creation

A side effect of the application of this patch is the creation of a new local extremum in the neighborhood of (x_1, y_1) as indicated by a red triangle on Fig. 2(b). The choice of a patch being not null over a relatively small region ($\mathcal{D}_1 \cup \mathcal{D}_2$) stems in a concentration of the energy needed to cancel difference d. This creates artifacts that in turn trigger the creation of new keypoints.

Table 1 shows that this is not an isolated or random phenomenon. There are almost as many new keypoints created as deleted ones. When giving the relationship between PSNR and keypoint removal rate, [2] not only does not count the new keypoints, that distort the given results, but it even does not mention their existence. In most of the cases, keypoints are not removed but indeed displaced.

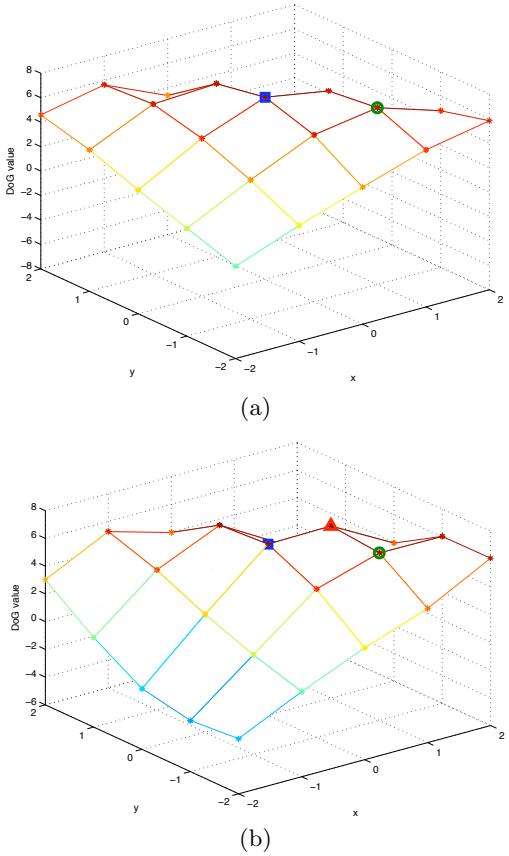
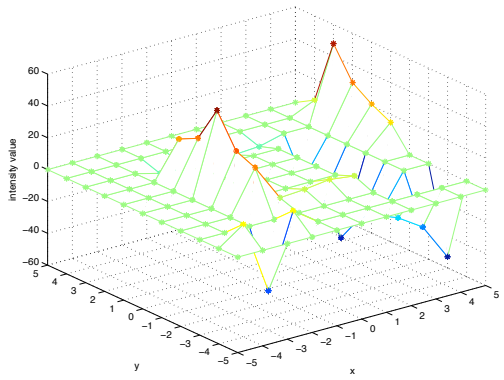


Figure 2: Effect of KPR-attack on the 5x5 neighborhood of a particular keypoint $(x, y, \sigma = 1.37)$: (a) original DoG values, (b) DoG values after attack.

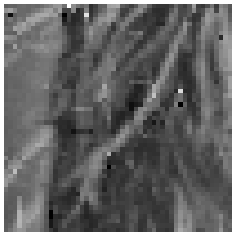
The next question is: what are the properties of new keypoints compared to those they replace? Fig. 4 seems to indicate that the new keypoints are very close to and at the same scale as the old ones. Of course the proximity of keypoints is relative to their scale. We measure this spatial proximity by computing the scale-normalized distance between original and new keypoint locations at same scale. The average distance over the 1,000 queries is 3.2, which proves the spatial proximity of new keypoints. Detailed results per octave for Lena are given in Table 2.

4.3 From Keypoints to Descriptors

The keypoints are used to compute descriptors over their neighborhood, called support region. It follows that if two keypoints are close to each other, their support regions are



(a)



(b)

Figure 3: The KPR-based attack for a particular keypoint $(x, y, \sigma = 1.37)$: (a) patch ϵ, (b) visual distortion on image induced by ϵ patch.

very likely to be similar, and therefore their descriptors also. Fig 5 shows this descriptor similarity when the location of the keypoint is artificially shifted by some pixels in the direction of the principal orientation of the original keypoint, which triggers the strongest changes in the computed descriptors. For high scale, keypoints must be moved farther away to significantly change the descriptor because their support region is larger. In the end, reducing the likelihood of match requires to displace keypoints very far away from their original location, and/or at a different scale. Consistently with the conclusions of the previous section, the KPR-attack of [2] fails to shift points, and this is the reason of its inefficiency against the CBIRS.

5. CONCLUSIONS

The conclusion is twofold. The attack proposed in [2] is not at all a threat for CBIRS, as removing keypoints triggers the creation of new and undesired ones that are easy to match. It may, however, impact other applications of SIFT: For instance, the KPR-attack might be efficient against some image authentication and robust hash schemes as also considered in [2].

Our paper does not prove that CBIRS based on SIFT are secure. There might be other dedicated attacks circumventing other bricks of the system like the description part [1]. Even the keypoint detection might be hacked, but attackers must be very careful about the creation of new keypoints.

6. REFERENCES

[1] T.-T. Do, E. Kijak, T. Furon, and L. Amsaleg. Challenging the security of content based image retrieval systems. In *Proc. MMSP*, 2010.

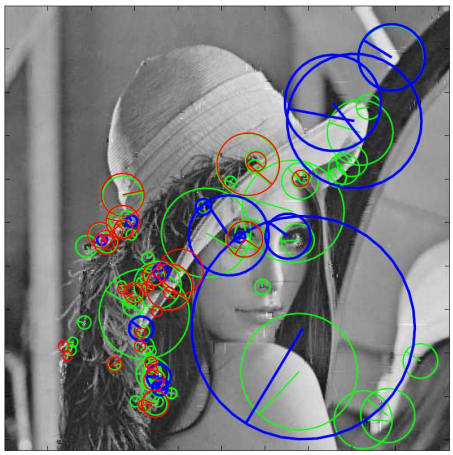


Figure 4: Representation of a subset of keypoints: the center is the keypoint location, the radius is proportional to the scale, and the dominant orientation is given by the represented radius. In blue: unchanged keypoints; in green: keypoints removed by KPR-attack; in red: created keypoints.

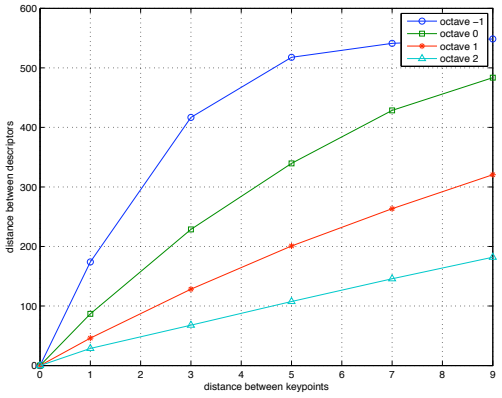


Figure 5: Euclidean distance between descriptors as a function of the distance in pixels between the keypoints, when displaced along the principal orientation of original keypoint, for different octaves on Lena image.

[2] C.-Y. Hsu, C.-S. Lu, and S.-C. Pei. Secure and robust SIFT. In *ACM Multimedia Conf.*, 2009.

[3] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford. Video copy detection: a comparative study. In *Proc. CIVR*, 2007.

[4] H. Lejsek, F. H. Ásmundsson, B. T. Jónsson, and L. Amsaleg. Nv-tree: An efficient disk-based index for approximate search in very large high-dimensional collections. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):869–883, 2009.

[5] D. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 60(2), 2004.

[6] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.