Deluding Image Recognition in SIFT-based CBIR Systems

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

Deluding Image Recognition in SIFT-based CBIR Systems

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

Content-Based Image Retrieval Systems used in forensics related contexts require very good image recognition capabilities. Therefore they often use the SIFT local-feature description scheme as its robustness against a large spectrum of image distortions has been assessed. In contrast, the *security* of SIFT is still largely unexplored. We show in this paper that it is possible to conceal images from the SIFT-based recognition process by designing very SIFT-specific attacks. The attacks that are successful in deluding the system remove keypoints and simultaneously forge new keypoints in the images to be concealed. This paper details several strategies enforcing image concealment. A copy-detection oriented experimental study using a database of 100,000 real images together with a state-of-art image search system shows these strategies are effective. This is a very serious threat against systems, endangering forensics investigations.

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 at the root of many forensics related applications such as the copyright enforcement of illegal multimedia contents or the dismantling of child pornography networks. CBIRS typically use a (large) database of images together with a sophisticated image description scheme having very good and robust recognition properties: query images can be matched with database images even if they are severely modified.

Recently, however, it is not the robustness of CBIRS that is still challenged, but its security [2, 3]. Robustness is the answer to general purpose attacks applied to images, such as cropping, color changes, rotations, filtering, ... Attacking the security of CBIRS is very different: attacks are instead very specific as they try to defeat the particular techniques used in one system. Pirates accumulate very fine-grain knowledge on these techniques and subsequently produce attacks exploiting their flaws. This is a very serious threat since we will soon face, for sure, elaborated security-challenging attacks such as [7].

This paper is a step toward understanding what is at stake when trying to delude the recognition power of CBIRS. We show that very specific attacks against the popular SIFT descriptors [6] succeed in severely endangering the image recognition process. This paper is structured as follows. Section 2 briefly discusses the properties of SIFT that matter when challenging their security. Section 3 details the two typical strategies to delude a system, respectively concealing an illegal image from identification and polluting the search results with false positives. Section 4 evaluates the impact of these strategies against a large scale database used together with a state-of-art CBIRS. It shows the strategies are effective in breaking the security of the SIFT description scheme. Section 5 concludes the paper and suggests some counter measures.

2. SIFT: KEYPOINT, PATCH, AND LOCAL DESCRIPTOR

SIFT [6] is probably one of the most popular description scheme extracting local features from images, and its excellent robustness has been assessed [4]. Given one image, SIFT first determines keypoints. Then it computes high dimensional vectors using the pixels surrounding each keypoint, taking scale into account. Image recognition is the result of the matching between the vectors of the query and the ones extracted from the database images. This matching is based on distances as a k-Nearest Neighbor (k-NN) process is ran for each query vector. Eventually, a similarity score is computed from these neighbors.

More formally, keypoint detection relies on local extrema of the Difference-of-Gaussian function $D(x, y, \sigma)$:

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,h\sigma) - G(x,y,\sigma)) \otimes I(x,y) \\ &= \Delta G_\sigma(x,y) \otimes I(x,y), \end{aligned} \quad (1)$$

MiFOR'10, October 29, 2010, Firenze, Italy.

where $\otimes$ is the 2D convolution operator, and $G(x, y, \sigma)$ is the kernel of the variable-scale Gaussian low-pass filter.

A keypoint is detected at location and scale $\mathbf{x} = (x, y, \sigma)^T$ if the following three conditions hold: (i) $D(\mathbf{x})$ is a local extrema over a neighborhood of $\mathbf{x}$; (ii) a sustainable contrast is present, i.e., $|D(\mathbf{x})| > C$ where C is a threshold hard-coded in the algorithm; (iii) the keypoint is not located on an edge, which can be detected by $\mathrm{tr}(\mathbf{H})^2 / \det(\mathbf{H}) < \tau$, with $\mathbf{H}$ the 2x2 Hessian matrix of $D(\mathbf{x})$.

Each keypoint is subsequently used to generate high dimensional vectors, so-called local descriptors. The local descriptor $V_{\mathbf{x}}$ is an illumination-invariant 128-bin histogram of gradient orientations around the keypoint $\mathbf{x} = (x, y, \sigma)^T$.

Even if the image is severely attacked, it is likely that a significant number of its local descriptors are still matching with the ones of the original image [6, 4]. There are several reasons for this. First, the robustness of SIFT comes in part from the keypoint extraction process which is very stable: a large proportion of identical keypoints are detected in similar, yet severely distorted, images. Second, these identical keypoints often define very similar visual patches (relative to their scales, however). This in turn results in the creation of local descriptors with very similar values along their dimensions, likely to match thanks to the k-NN search paradigm. Last, the robustness of SIFT also comes from the large number of local descriptors describing one image, typically around one thousand for a 512x512 pixels image. Even if a large portion of the keypoints differ between an original image and its distorted copy, even if the visual distortions are such that it additionally impacts the local descriptors themselves, these two images need only to share a few tenth of actual matches for assessing recognition. In fact, the non-matching local descriptors are likely to vote for images randomly distributed over the whole database, failing to significantly increase the score of a particular image.

3. STRATEGIES CHALLENGING SIFT

The performance of a CBIRS is mainly related to the ability of the description scheme in use to match images despite distortions. Without loss of generality, the search process eventually builds a ranked result list of images, such that highly ranked images are more likely to be true positive matches. True positives tend to have a particularly high score compared to the other images in the list. Two strategies can therefore be designed to delude the recognition of a system. First, it is possible to attack the image to be concealed such that its score gets dramatically reduced. Second, it is possible to attack the image by introducing in the picture visual elements that often match with other images from the database, such that the image to be concealed gets ranked far in the result list. It is of course possible to combine these two strategies.

This section describes in details several ways to instantiate these strategies. It visually illustrates their impact using the well known Lena image. For fairness, it also shows their impact when applying the strategies to the set of 1,000 real images used in the performance evaluation section of this paper (see Section 4). Note we computed the local SIFT descriptors using the open-source SIFT-VLFeat code by Vedaldi [8]. We did several experiments to get SIFT-VLFeat 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 (C) and edge-thresh=12, and 3 intervals per octave.

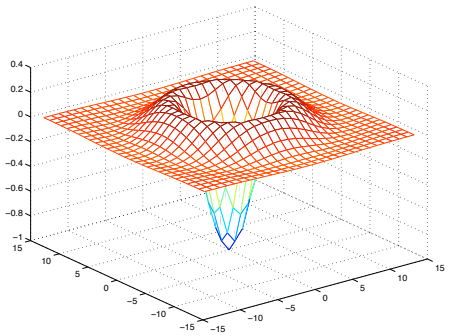

Figure 1: A RMD patch has been applied at scale $\sigma = 2.54$ with $\delta = -1$. The z-axis represents the difference $D'(x+u, y+v, \sigma) - D(x+u, y+v, \sigma)$. If the patch succeeds to lower the targeted DoG coefficient of the amount δ, it also impacts the DoG values around.

3.1 Reducing Scores

Concealing an image from recognition by reducing its score translates into reducing the number of local descriptors that match. This can obviously be achieved by eliminating some of its keypoints, but also by modifying the location and/or scale of keypoints in order to generate different values of the local descriptors so as to move them far away from their original position in the high dimensional feature space.

3.1.1 Keypoint Removal

As explained in section 2, keypoints are detected when local extremum are found in DoG with enough contrast and away from edges. Avoiding the detection of a keypoint at $\mathbf{x} = (x, y, \sigma)^T$ is possible by applying on the image a visual patch of a certain size centered at (x, y) such that at least one of the three conditions above does not hold anymore. Patching images introduces visual artifacts, and, thus, a side effect of keypoint removal is the creation of new, undesired keypoints. We study in the following two keypoint removal strategies that are extreme in the sense that one is designed to minimize the *local* distortion on images, ignoring any potential keypoint creation, while the other takes little care of the distortion but rather eliminates as many keypoints as possible while lowering the number of creations.

Removal with Minimum Local Distortion

For the keypoint $\mathbf{x} = (x, y, \sigma)^T$, this approach determines a patch ϵ to apply on the image such that it minimizes the *local* distortion. Since, at scale σ, the Difference of Gaussian kernel ΔG_σ has a limited support $\mathcal{S}_\sigma$ in the spatial domain, ϵ defined over $(x, y) + \mathcal{S}_\sigma$ modify the quantity $D(x, y, \sigma)$ of a given amount δ. In other word, for (u, v) in the neighborhood of (x, y), the image is modified in $I'(u, v) = I(u, v) + \epsilon(u, v)$ so that $D'(\mathbf{x}) = D(\mathbf{x}) + \delta$. The patch should be of minimal Euclidean norm to reduce the perceptual degradation. This obviously resorts to an optimization under constraint:

$$\epsilon = \arg \min_{\epsilon : D'(\mathbf{x}) = D(\mathbf{x}) + \delta} \|\epsilon\|^2 \qquad (2)$$

(a) (b) (c) (d)

Figure 2: Visual distortions caused by: (a) RMD with $C = 4$, $\delta^+ = 4$ for keypoints at different scales (see on the lips and the shadow under the left eye), (b) GS+LS7, (c) FMD with $C = 4$, $\delta^- = 3$, and (d) GS+LS7+FMD.

The constraint being affine and the function to be minimized being convex, a simple Lagrangian resolution yields that

$$\epsilon = \frac{\delta}{\|\Delta G_\sigma\|^2} t_{(x,y)}(\Delta G_\sigma),$$

where $t_{(x,y)}$ is the 2D translation operator of a shift (x, y). This patch controls $D(\mathbf{x})$, however, as illustrated in Fig. 1, it also modifies the DoG values in the neighborhood such as $D(x + u, y + v, \sigma)$ with $(u, v) \in [-6h\sigma\sqrt{2}, 6h\sigma\sqrt{2}]^2$ where h is a constant factor in scale space, chosen as $h = 2^{\frac{1}{s}}$ with s the number of intervals in an octave.

Let C be the fixed contrast threshold. We control the Removal with Minimum local Distortion attack (RMD) power by targeting a limited number of keypoints to be erased. We introduce a value $\delta^+ > 0$ that defines the subset $\mathcal{E}_{\delta^+} = \{\mathbf{x} : C < |D(\mathbf{x})| < C + \delta^+\}$. Erasing keypoints in $\mathcal{E}_{\delta^+}$ means that we decrease $|D(\mathbf{x})|$ by an amount $|\delta|$ such that its new value is below the threshold C: $|\delta| = |D(\mathbf{x})| - C$. Obviously, when δ^+ grows, we deal with a bigger subset $\mathcal{E}_{\delta^+}$, and more keypoints are removed. The most visible artifacts are caused by the removal of keypoints at higher scales, mostly because the patches have a bigger support and a stronger amplitude, as shown in Figure 2(a). Note with RMD, it is the local distortion around each keypoint that is minimized, not the global distortion on the whole image.

However, as shown in Table 1, the RMD attack triggers the creation of new keypoints. Table 1 summarizes the results when comparing keypoints between the attacked and the original images. This is also illustrated on Figure 3.

Maximum Removal, Minimum Creation

Local extrema of the DoG correspond to points located on significant discontinuities in the image. A straightforward way to avoid their detection consists in smoothing the image. Performing a smoothing on the whole image reduces the number of keypoints while minimizing the creation of new ones, as it does not introduce strong discontinuities. Experiments show this global smoothing is quite effective when using a Gaussian kernel of small variance ($\sigma = 1.3$). A greater σ would in turn remove more keypoints, but the quality of the resulting image would be very bad. This global smoothing strategy is referred to as the GS strategy.

To increase the number of removed keypoints, a second step of smoothing can subsequently be applied, this time on a local basis. After having applied GS to one image

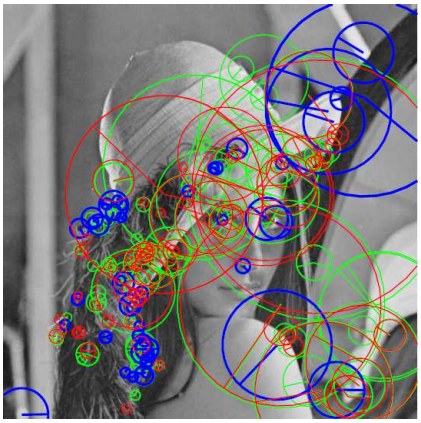

Figure 3: Keypoints removal with the RMD attack on Lena image with $C = 4$ and $\delta^+ = 7$ illustrated on a subset of keypoints: unchanged (blue), deleted (green), and new (red) keypoints.

to decrease the number of keypoints, a local smoothing (LS) phase can be ran on each remaining keypoint: it replaces the $n \times n$ region around the current keypoint by its smoothed version with a gaussian kernel whose variance equals the keypoint scale, and checks whether this keypoint is still detected or not. This is in fact performed iteratively using regions of growing sizes ($n = \{1, 3, 5, 7\}$), until the keypoint is no longer detected. Having n taking values larger than 7 introduces too severe distortions in the images. Therefore, if the keypoint is still detected when $n = 7$, then the LS phase is aborted for that keypoint and the keypoint is left in its original state. This corresponds usually to keypoints with large scales. The strength of the LS attacks can be controlled by the maximum value we allow n to take in $\{1, 3, 5, 7\}$: the strongest LS attack, noted LS7, is when n can take all values, while with the (less strong) LS5 attack n varies only within $\{1, 3, 5\}$. The visual distortions introduced by LS7 are shown in Figure 2(b). Note one image can be attacked by applying GS only, or GS+LS.

As shown in Table 1 and Figure 4, both global and local smoothing also create new keypoints. Applying LS after GS results in a smaller number of created keypoints, as keypoints created by GS may be subsequently removed by LS. The number of created keypoints, however, is by far smaller than when using the RMD attack.

Table 1: Number of deleted and created keypoints by RMD with $\delta^+ = 7$, GS, GS+LS7, and GS+LS7+FMD.

Image	Attack	# KP deleted	% KP deleted	#KP created	% KP created	# KP after attack	PSNR in dB
Lena (1218 keypoints)	RMD	1102	90.48	888	72.91	1004	27.78
	GS	1112	91.30	339	27.83	445	31.17
	GS+LS7	1172	96.22	270	22.17	316	30.31
	GS+LS7+FMD	1174	96.39	383	31.44	427	30.12
1,000 imgs (1026 keypoints, average)	RMD	795	77.49	546	53.22	777	30.70
	GS	903	88.01	395	38.50	518	29.17
	GS+LS7	967	94.25	321	31.29	380	28.41
	GS+LS7+FMD	967	94.25	389	37.91	448	28.23

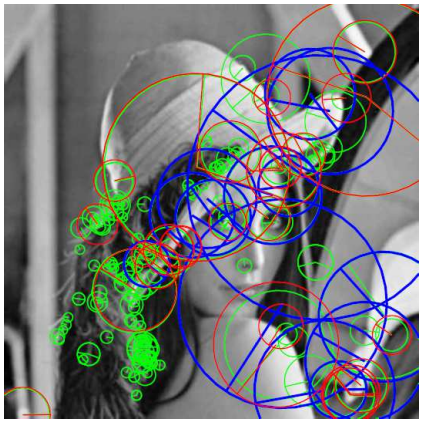

Figure 4: Smoothing Lena with GS+LS7 illustrated on a subset of keypoints: unchanged (blue), deleted (green), and new (red) keypoints.

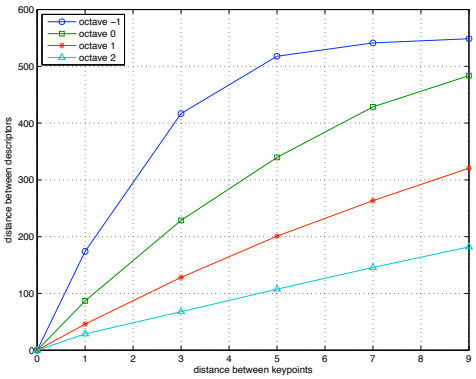

Figure 5: Euclidean distance between descriptors vs. distance between keypoints (Lena image).

3.1.2 Shifting Keypoints in Space and/or Scale

Removing keypoints to eventually reduce the scores at matching time is one option. Another option is to shift keypoints in the images, either in their (x, y) location in the image or in their scale. Shifting one keypoint in the image space also shifts the patch of pixels used for the descriptor computation, and hence is likely to change the final high-dimensional descriptor. Shifting the scale of one keypoint has similar effects: in this case, however, the patch does not move but its radius gets increased or decreased, which, in turn, gives different pixels for building the final descriptor. Overall, this strategy does not remove keypoints but reduces the likelihood with which they will match with the original, non attacked, image. Intuitively, the shift must be large enough in order for the final descriptor to be sufficiently different and to fail matching.

We have not yet implemented a fully automatic version of this strategy but it is possible to gain knowledge on its effectiveness by artificially shifting keypoints in space and/or scale, thanks to the open source implementation of the SIFT-VLFeat [8]. We therefore conducted two small experiments in order to understand how large must be the shift in space or in scale to significantly reduce the likelihood of matching. We carefully tracked the whole descriptor creation process for all keypoints of the Lena image. We then replayed the creation for these keypoints, but we artificially shifted each keypoint away from its original location in space (or in scale) and measured the euclidean distance between the descriptor associated with that shifted keypoint and its original version. Figure 5 shows the result of this experiment when considering spatial shifts while Figure 6 is for scale shifts. Figure 5 shows that shifting keypoints in the image indeed increases the distances in the high-dimensional space. Scale matters, however. Overall, this reduces the probability for that shifted descriptor to be part of the k-NN list. A similar behavior is observed when shifting the scales of keypoints.

Table 2 gives indications on the locations and scales of the keypoints created as a collateral effect of applying RMD, GS+LS (and GS+LS+FMD, see next section). It focuses on the new keypoints belonging to the first two octaves (-1 and 0), as it is where the majority of creations take place. For each such new keypoint, we computed its distance to the nearest original keypoint in the same octave. Averaging for all created keypoints gives the "avg distance" line in Table 2. That table also gives the average scale factor between new keypoints and their nearest original keypoints, regardless of the octave. It can be seen that new keypoints created by LS+GS are farther away both in location and especially in scale than those created by RMD. By using these values together with Figures 5 and 6, it is possible to get a rough idea about the general evolution of distances in the feature space between the new and original descriptors. This corroborates the fact that LS+GS attacks are stronger than RMD attacks.

3.2 Triggering False Positives

As mentioned in the introduction, the other strategy is to deliberately create new keypoints so that their descriptors will match with wrong images in the database, increasing their scores. Hopefully, querying with the attacked image will bring at the top of the result list some other images,

Table 2: Properties of the new keypoints: average distance per octave and scale factor between new keypoints and the nearest original keypoints. The last column only shows the creation of keypoints due to FMD.

Image	Lena						1,000 imgs					
Attack	RMD, $\delta^+ = 7$		GS+LS7		GS+LS7+FMD		RMD, $\delta^+ = 7$		GS+LS7		GS+LS7+FMD	
Octave	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0
# KP created	613	188	11	147	68	43	329	141	28	196	36	32
avg distance	3.1	9.1	4.1	7.8	36.3	34.0	4.1	8.7	4.8	10.7	34.6	36.3
avg scale factor	0.21	0.62	0.47	0.80	0.56	0.79	0.26	0.55	0.45	0.77	0.51	0.75

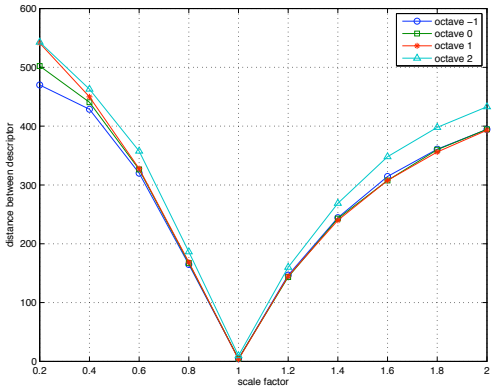

Figure 6: Euclidean distance between descriptors vs. scale shift of a given factor (Lena image).

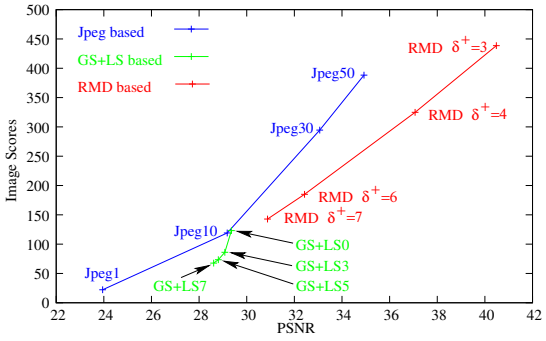

Figure 7: Average score vs. PSNR (dB). JPEG-, GS+LS and RMD based attacks.

while the original one may either be further away in the list or not in the list at all. This also raises false alarms spoiling the trust users can have on the system.

The strategy we use for Forging new keypoints with Minimum local Distortion (FMD) is symmetric to the RMD attack. In this case, we address the local extrema in the subset $\mathcal{F}_{\delta^-} = \{\mathbf{x} : C - \delta^- < |D(\mathbf{x})| < C\}$, and add patches that strengthen the contrast in the neighborhood of keypoints. This also reduces the gap between the absolute values of the first and second eigenvalues, such that the condition on the Hessian matrix gets verified most of the times. The new keypoints are easy to add, especially in the first octave, resulting in salt and pepper noise artifacts (Figure 2(c)).

To be created with FMD, new keypoints must meet two conditions: (i) belong to the two first octaves, such that the introduced distortion is small, (ii) be located relatively far away from existing keypoints such that the resulting descriptors are rather different from the existing ones. We thus create keypoints where their neighborhood of size 8 pixels is free from original keypoints. Table 2 shows their average distance from the nearest original keypoint. As expected, this distance is sufficiently important to trigger descriptors that will not match with the original ones.

4. LARGE SCALE EXPERIMENTS

4.1 Dataset, Queries, Experimental Protocol

We evaluate the efficiency of the attacks using a large scale image collection and a real CBIRS. 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 high-dimensional indexing scheme [5]. The NV-Tree runs approximate k-NN queries and has been specifically designed to index large collections of local descriptors. We then randomly picked 1,000 of these images and ran 9 "security" attacks on them. For comparison, we also applied 49 standard Stirmark "robustness" attacks (rotations, crops, filters, scalings, ...). Overall, there are 58,000 queries distributed in 58 families. This experimental protocol clearly targets a copy detection scenario. For all queries, we record the scores of the 100 best matching images.

4.2 Results

The first experiment illustrates the efficiency of keypoint removal. We average the score of the 1,000 original images when searched with 4 RMD attacks and 4 GS+LS attacks, varying the strength of the attacks. We also collect the average PSNR observed on the attacked images for each family of attack. The results of this experiment are on Figure 7, which additionally depicts scores and PSNR for JPEG attacks with varying quality factors, for comparison. As expected, both the score and the PSNR drop down as the attack strength increases (through increasing parameters n or δ^+). It also confirms that the RMD attack yields a better PNSR than GS+LS, as it was designed to introduce the minimum distortion. RMD is less efficient than GS+LS at lowering the matches, as it creates more descriptors that are not so different from their original counterpart.

Figure 8 illustrates the result of the full experiment, focusing on the 9 "security" attacks. The results for the Stirmark "robustness" attacks are skipped since none concealed copies; they were used as sanity checks, however. Along the X-axis, the attacks differ by the keypoint removal process with varying parameters as described in section 3 and whether the FMD attack is turned off (left side) or on (right side). The Y-axis of the figure shows, for each attack, the average scores (over the 1,000 queries) of the original images that are expected to match with the attacked copies. It also shows the scores of the first, second and third best match-

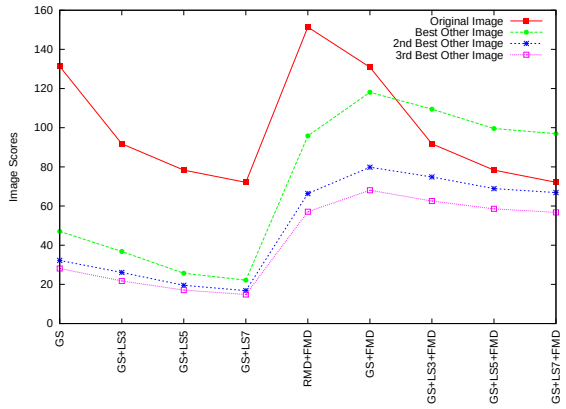

Figure 8: Image scores in realistic settings. X-axis: 9 selected "security" attacks. Y-axis: for each family of attack, the average scores over 1,000 queries of the original images and of the three other best ones.

ing images that are different from the original images. When the image having rank #1 in the result list is the original image, then the system succeeded in recognizing that image. When the image at rank #1 *is not* the original image, then the system failed. For comparison, the score of the original images when searched with themselves is around 1,000 and the second best is below 100.

4.3 Analysis

Figure 8 shows that solely removing keypoints (the 4 attacks on the left) fails at deluding the system. There are mainly two reasons. First, the gap in scores between the original image and the best non-matching image is large, and therefore, at least 90% of the keypoints should be removed to shrink that gap, which is impossible in practice due to the severe visual distortion this would cause. Second, keypoint removal creates new keypoints in their vicinity and it turns out the corresponding vectors match with the original ones.

When FMD is turned on then things are changing. With attack GS+FMD, although the original image is found, the score of the best other image jumps and get much closer. A detailed examination of the matching shows that the new forced keypoints are created at small scales and tend to match with images from the database that have repetitive visual patterns such as bricks, small windows on large facades, tiles. Interestingly, despite the size of the collection, few images have such patterns (see Fig. 9 for an example) and therefore they concentrate the votes when scoring.

Increasing the strength of these attacks by adding local smoothing succeed in deluding recognition. With the last three attacks on the right of the Figure 8, not only few other images concentrate matches (thanks to FMD) but the number of keypoint in the attacked image drops (thanks to GS+LS7), reducing the number of true matches. It is essential to note the attacked image is not concealed (as its unmodified copy has rank #2), but it gets "hidden" behind another image that better matches. This is a key result.

Overall, these experiments are a proof of concept that "security" attacks can conceal an image, at the cost of a distortion of around 30 dB in PSNR (see Table 1). A "robustness" attack, such as a JPEG compression with $Q = 1$, achieves the same goal but with a PSNR of 23.68dB (see Figure 7).

Figure 9: Images of the database often ranked #1 when the attack succeeds.

Two major lessons can be drawn from this study. First, solely removing keypoints is not sufficient to delude a system in real settings. Second, the forgery of new keypoints is necessary, best if they match with a restricted set of images from the database. These lessons open several interesting issues. One is to have the pirate creating artificial images designed to concentrate votes before inserting them into the database. This eventually "pollutes" the result list as these images will get ranked high, hiding the actual matches.

5. CONCLUSION

In this paper, we play the role of a pirate wishing to delude a CBIRS, in the sense that he wants to conceal from recognition a copy of a given image in the database. We show that "security" attacks dedicated to a given CBIR technique are more efficient than 'robustness' attack. In our future works, we will play the role of the designer, investigating possible counter-measures. With respect to the attacks presented in this paper, a possible counter-measure is to use the typical post-processing steps removing the false positives from the result list thanks to geometrical verification [1] for example.

6. REFERENCES

[1] M. Douze, H. Jégou, and C. Schmid. An image-based approach to video copy detection with spatio-temporal post-filtering. *IEEE Trans. on Multimedia*, 2010.

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

[3] E. Kijak, T. Furon, and L. Amsaleg. Challenging the Security of CBIR Systems. Research Report RR-7153, INRIA, 2009.

[4] 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.

[5] H. Lejsek, F. H. Amundsson, B. T. Jonsson, and L. Amsaleg. Nvtree: An efficient disk-based index for approximate search in very large high-dimensional collections. *IEEE TPAMI*, 31(5), 2009.

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

[7] S. Smitelli. Fun with youtube's audio content id system. `http://www.csh.rit.edu/~parallax/`.

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