Visual Security Evaluation Based on SIFT Object Recognition

Stefan Jenisch, Andreas Uhl

Visual Security Evaluation Based on SIFT Object Recognition

Stefan Jenisch and Andreas Uhl

Multimedia Signal Processing and Security Lab (WaveLab)
Department of Computer Sciences, University of Salzburg
andreas.uhl@sbg.ac.at

Abstract. The paper presents a metric for visual security evaluation of encrypted images based on object recognition using the Scale Invariant Feature Transform (SIFT). The metrics' behavior is demonstrated using three different encryption methods and its performance is compared to that of the PSNR, SSIM and Local Feature Based Visual Security Metric (LFBVSM). Superior correspondance to human perception and better responsiveness to subtle changes in visual security are observed for the new metric.

1 Introduction

Today a number of (format compliant) image encryption techniques exist which allow the encrypted content to be decoded and viewed. To determine the level of security offered by these techniques it is not enough to simply evaluate the cryptographic strength of the encryption cipher used.

For some methods the decoded encrypted image is a low quality version of the original image and certain image features can still be recognised. So beside evaluating the encryption cipher also the *visual security* of the result has to be assessed. In this context we need to deal with the remaining image quality left behind by the encryption process and the recognizability and intelligibility of the encrypted image content.

In order to be able to discuss the exact notion of visual security, we need to distinguish distinct application scenarios of media encryption schemes [1]:

Confidentiality Encryption: Means MP security (mes- sage privacy). The formal notion is that if a system is MP- secure an attacker cannot efficiently compute any property of the plain text from the cipher text. This can only be achieved by the conventional encryption approach, i.e. applying a cryptographically strong cipher to compressed (redundancy-free) image data.

Content Confidentiality: Is a relaxation of confidential encryption. Side channel information may be reconstructed or left in plaintext, e.g. header information, packet length, but the actual visual content must be secure in the sense that the image content must not be intelligible / discernible.

Sufficient Encryption: Means we do not require full security, just enough security to prevent abuse of the data. The content must not be consumable due to high distortion (e.g. for DRM systems) by destroying visual quality to a degree which prevents a pleasant viewing experience or destroys the commercial value. This implicitly refers to message quality security (MQ), which requires that an adversary cannot reconstruct a higher quality version of the encrypted material than specified for the application scenario.

Given these different application scenarios it is clear that depending on the goal, a security metric has to fulfill dif- ferent roles. For example, under the assumption of sufficient encryption a given security metric would have to evaluate which quality is low enough to prevent a pleasant viewing experience.

When it comes to content confidentiality the question of quality is no longer applicable. Content confidentiality requires that image content must not be identified by human or automated recongnition. This requirement also has to be maintained for any part of the image. Image metrics, in general, do not deal with such questions but rate the overall image quality, the question of intelligibility is usually not covered at all. Thus, it seems to be clear that a general purpose metric covering all application scenarios is probably very hard or impossible to design.

Additionally we have to face the fact that different encryption methods introduce different kind of distortions. While some methods shift and morph the images (i.e. chaotic encryption which is mainly based on permutations) others introduce noise and noise like patterns. An ideal metric for assessment of visual security has to be able to deal with those different kind of distortions.

To evaluate the visual security of an encrypted image in an objective manner, often the Peak Signal to Noise Ratio (PSNR) or the Structural Similarity (SSIM) Index are used. Despite the fact that both originally have been developed for image quality assessment, they have also been used for the assessment of encrypted images [2,3,4,5].

Also several attempts have been made to develop a metric specifically for the task of visual security assessment. One popular example is the Local Feature Based Visual Security Metric (LFBVSM [6]), which compares corresponding image regions of the cipher and plain images by their luminance and contour information to evaluate the visual security of an encrypted image. Also further dedicated metrics for visual security evaluation have been proposed (e.g. [7,8,9]).

While there exist particular image encryption techniques for which PSNR, SSIM, and LFBVSM do a reasonable job to rate the visual security of a ciphered image, for many encryption methods these metrics tend to have troubles in the correct assessment of visual security in correspondence to visual perception (as we shall see in the experiments).

Since most of these metrics compare the plain and the cipher images pixel by pixel or region by region (fundamental principles of the Human Vision System (HVS) in terms of luminance and edge perception are considered) a warped image may still be recognisable while the metric rates the image as secure due to large dissimilarities in terms of pixel or local region differences. Also, noise patterns tend to decrease the score rater quickly but leave the content of the image still intelligible. Thus, answering the question if an encryption of this type results in a *content confidential* image, i.e. an image without any intelligible content, can become quite challenging with those metrics.

In this paper, we aim to apply object recognition methods to design a metric for visual security assessment in order to tackle the issue of content recognition and intelligibility in a more appropriate manner. The basic idea of the metric presented is to compare the recognizability of objects found in a reference and a cipher image instead of measuring image quality as such.

In particular, we propose to employ the Scale Invariant Feature Transform (SIFT) [10] for object recognition in ciphered images. Therefore, the metric is termed "Scale Invariant Feature Transform Similarity Score" (SIFTSS).

The paper is divided into four parts. First a description of the SIFTSS is given, followed by a description of the encryption methods used to test the metrics performance. Then the results of the experiments are presented and finally they are discussed in the last section of the paper.

2 SIFT Similarity Score

The SIFT algorithm derives a set of key-points for each image. Each key-point is associated with a descriptor vector (edge histograms). The images are compared using these key-points, i.e. all key-points of the target (ciphered) image are compared to the key-points found in the reference (original) image. The matching process compares the Euclidian distances between descriptor vectors of the reference key-points and the target key-points. A search for the minimum Euclidian distance between their descriptor vectors is carried out.

To improve matching performance, a validity check is performed. The minimum distance found is multiplied by the value of 1.5 and again compared to the set of distances. If the multiplied distance is still smaller than all other distances, the key-point pair involved is considered a match. The check ensures that the match of the key-points is distinctive or dominant in comparison to all other possible matches.

For the implementation of the metric the VLFeat.org [1] implementation of the SIFT algorithm was used. The matching process returns an array of matching image key-points along with their corresponding Euclidian distances, measured between their descriptor vectors. The number of matched key-points as well as the average Euclidian distance of the edge histograms is used to derive a matching score.

In the formula

$$\mathrm{SIFTSS}(A,B) = \left(\frac{\dim(m)}{\min(n_A, n_B)}\right)^{\dfrac{\mu_m}{|m|_2}} \tag{1}$$

the calculation of the SIFTSS between the images A and B is shown, where m is a vector containing a list of the Euclidian distances of the matched key-points between A and B, and n_A,n_B is the total number of key-points found in

[1] `http://www.vlfeat.org`

each of the two images. The number of matched key-points $dim(m)$ is divided by the maximum number of possible matches, and thus mapped into the interval $[0 : 1]$. The term is then taken to the power of the average Euclidian distance μ_m divided by the L_2-Norm of m. This also maps the exponent into the interval $[0 : 1]$. Consequently, while an increasing number of matched key-points increases the score, large Euclidian distances decrease it. Since the SIFT matching process is not commutative and returns different values when matching `imageA` to `imageB` than when matching `imageB` to `imageA` the algorithm calculates both directions and averages the result to restore symmetry.

In Listing 1.1 a pseudo code for calculating the SIFTSS is shown.

Listing 1.1. The calculation of the SIFTSS

```
[matchScoresA] = SIFTmatch(imageA, imageB);
[matchScoresB] = SIFTmatch(imageB, imageA);

matchCountA = length(matchScoresA);
matchCountB = length(matchScoresB);

if (matchCountA == 0 || matchCountB == 0)
  return 0
else
  normA = Norm(matchScoresA);
  normB = Norm(matchScoresB);
  if (normA == 0 || normB == 0) then
    return 0
  else
    normA = Norm(matchScoresA);
    normB = Norm(matchScoresB);
    if [normA == 0 || normB == 0] then;
      return 1;
        else
      meanEuclidianDistanceA = mean(matchScoresA/normA);
      meanEuclidianDistanceB = mean(matchScoresB/normB);

      keyPointsOfA = getNumberOfKeypoints(imageA);
      keyPointsOfB = getNumberOfKeypoints(imageB);

      matchScoreA = matchCountA / min(keyPointsOfA, keyPointsOfB);
      matchScoreA = power(matchScoreA, meanEuclidianDistanceA);

      matchScoreB = matchCountB / min(keyPointsOfA, keyPointsOfB);
      matchScoreB = power(matchScoreB, meanEuclidianDistanceB);

      matchScore = (matchScoreA + matchScoreB)/2

      return matchScore;

    end
  end
end
```

The range of SIFTSS values is between 0 and 1 where scores close to 0 signify a better visual security and 1 indicates identical images.

3 Experimental Settings

To give an overview of the metrics performance three case studies using different encryption methods have been selected and will be briefly described below. To establish a standard of comparison these cases are also evaluated using the PSNR, SSIM and LFBVSM. For the experiments the images of the *Kodak Lossless True Color Image Suite*[2] where cropped into a square format (which is required for Arnolds's Cat Map encryption), and scaled down to 150 × 150 pixels. Metrics results are averaged for this data set, all images have been encrypted using individual random encryption keys.

[2] `http://r0k.us/graphics/kodak/`

The first encryption approach uses Arnold's Cat Map, a chaotic map, for image encryption [11,12]. This type of encryption uses warp and shift operations for rendering the image unintelligible. The map works on an image of size NxN and is defined by the formula

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = C(x_k, y_k) = \begin{pmatrix} 1 & b \\ a & ab+1 \end{pmatrix} * \begin{pmatrix} x_k \\ y_k \end{pmatrix} mod\ N \tag{2}$$

where (x_k, y_k) satisfying $0 \leq x_k, y_k < N$ are the positions of the pixels in the original (square) image area while (x_{k+1}, y_{k+1}) is the position of the pixel in the target image, k is the number of the current iteration and $0 \leq a, b < N$, $a, b \in \mathbb{N}$ are the control parameters of the function used as key.

In each iteration the image is warped, cut and transformed back into its squared shape rendering the image more and more random. The operation has the property of a torus restoring the image after a discrete number of iterations. A visual explanation of a single iteration step is shown in Figure 1.

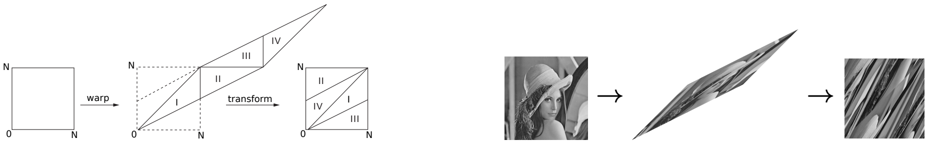

Fig. 1. Arnold's cat map in pictures

The iteration stages evaluated in the experiment were 0, 1, 3, 132, 155, 157, 200, 211, 250, 275, 299 and 300. In Figure 2 one of the encrypted images can be seen in all investigated iteration stages.

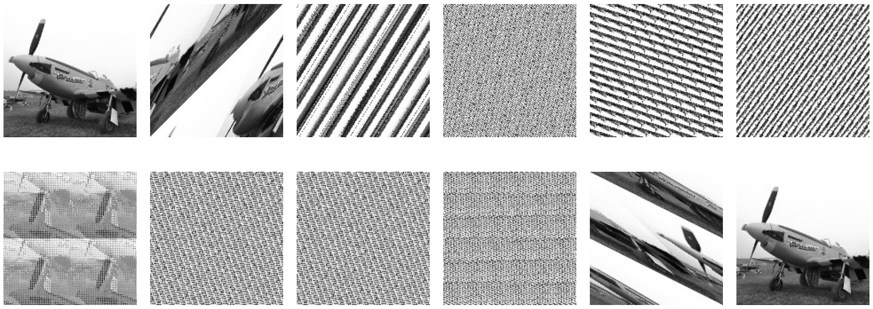

Fig. 2. Image transformed with Arnold's cat map (iterations ordered from left to right and top to bottom)

The second encryption method is integrated into the JPEG XR compression standard [13]. It is suggested to encrypt the DC coefficients using Random Level

Shifts (RLS), i.e. to alter the value of the DC coefficients by adding or subtracting random numbers which are derived from a key. To increase the impact of the encryption method on visual security we have recently suggested to apply RLS to all coefficients of a transform block, not only to its DC coefficient [1].

This encryption method introduces noise into the image and the impact on image perception can be seamlessly adjusted from low to high by setting the maximum allowed shift value accordingly. In Figure 3 sample images for increasing maximum shift values can be seen.

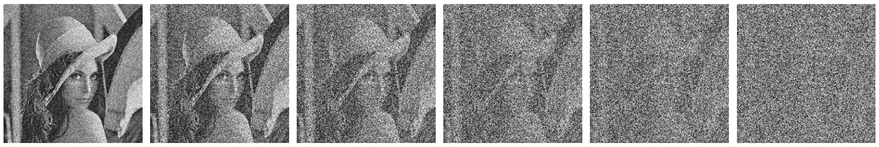

Fig. 3. RLS encryption using increasing maximum shift values (values from left to right: 80, 160, 280, 480, 800, 2000)

The third encryption method used for SIFTSS evaluation is the permutation of the coefficients' scan order in a JPEG XR code stream. This method has first been discussed for the JPEG standard [14] and was later proposed to be used for encrypting the LP frequency band of JPEG XR encoded images [15].

In JPEG XR the coefficients are grouped into three frequency bands, the DC-, the Lowpass- (LP) and Highpass (HP) band. In the experiment only the coefficients of the LP and HP band are subject to the permutation process [1]. Swapping of coefficients across frequency bands is not carried out.

As a result there are six possible encryption settings. For each of the two storage modes (spatial and frequency storage mode), the encryption of the LP, of the HP and of both frequency bands can be selected. Sample images for each encryption mode can be found in Figure 4.

To establish a subjective order in terms of visual security the images in Figure 4 are ordered from left (low security) to right (high security). Corresponding settings are: Spatial store mode and HP band encryption, Frequency store mode and HP band encryption, spatial store mode and LP band encryption, spatial store mode LP+HP band encryption, frequency store mode LP band encryption and frequency store mode LP+HP band encryption.

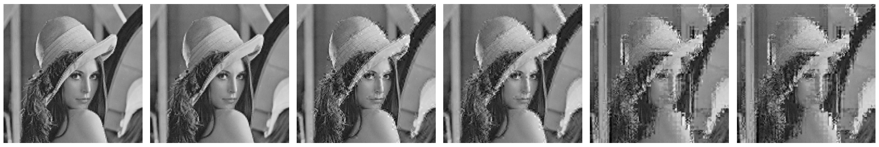

Fig. 4. Lena image scrambling using coefficient scan order permutation.

4 Experimental Results

The objective image metrics return values for the first experiment using Arnold's Cat Map can be seen in Figure 5.

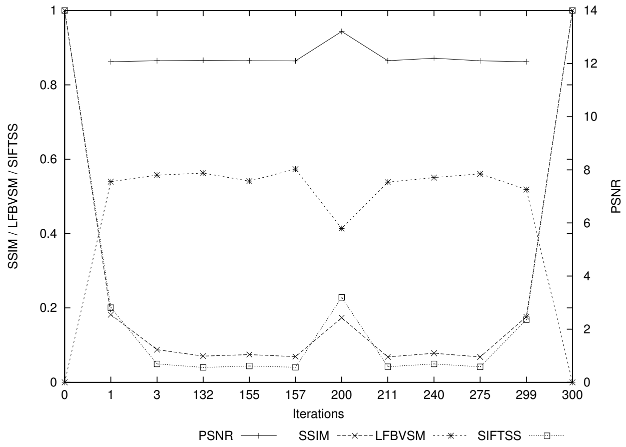

Fig. 5. Averaged return values for Arnold's Cat Map encrypted Kodak Image Database images.

The desired behavior of a metric evaluating this set of images would be an indication of content intelligibility for iteration stages 1, 200 and 299 and a contraindication for all others (stages 0 and 300 set aside). Subjectively, these iteration stages do show some resemblance of the original image (see Figure 2).

The SIFTSS indeed shows the desired behaviour for the suspected iteration stages 1, 200 and 299. Interestingly, also the SSIM indicates intelligibility for these stages. The PSNR and the LFBVSM do not show exactly the desired behavior. Notice that no values for iteration stage 0 and 300 are plotted in Figure 5 since the metric returns ∞ for identical images. Both metrics indicate some intelligibility for iteration stage 200 when comparing the return values for this stage with iteration stages 3 to 275. However, they fail to indicate it for iteration stage 1 and 299 which show similar values as for all other iteration stages.

The reason why iteration stage 1 and 299 is handled well by the SSIM may be explained with the fact that most pixels are still neighbouring each other in these stages. Certain image regions are not moved far from their original position (bottom left and top right corner of the image) and additionally certain regions show the same structure after the operation (e.g. an area showing clear blue sky is replaced with clear blue sky during warping). The SSIM metric uses a sliding window approach assigning each pixel of the image (the pixel aligned to the center pixel of the window) a similarity score derived from the area covered from the window. Also the metric uses mainly mean luminance and variance to describe the local structure information but ignores edge orientation and magnitude information which becomes disturbed during the warp operation. The LFBVSM metric however, which also compares image regions, uses edge orientation and magnitiude information contrasting to SSIM. This is probably the reason why the LFBVSM does not indicate intelligibility for those two iteration stages and SSIM does.

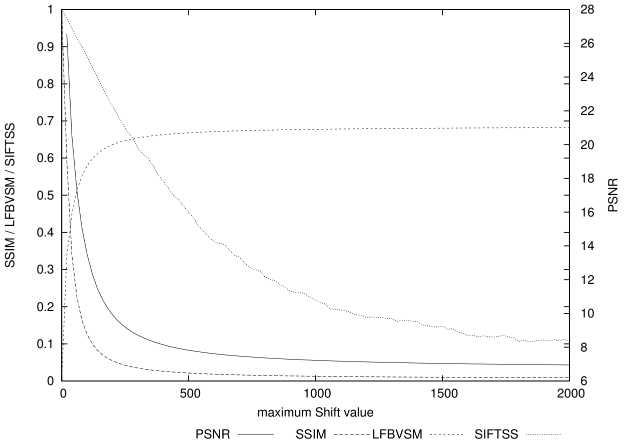

Fig. 6. Averaged return values for RLS encrypted Kodak Image Database images.

The objective image metrics averaged return values for the RLS encrypted Kodak Image Database images can be seen in Figure 6. The PSNR, SSIM and LFBVSM values exhibit a steep drop when increasing the maximum shift level from 0 to 300. After reaching a maximum shift value of 300 their slopes flatten out. As can be seen in Figure 3, this behaviour does not correspond well with visual perception where a gradual worsening of image quality and content intelligibility is observed across the entire range of considered shift values.

The SIFTSS on the other hand shows a much less steep drop in its return values when increasing the maximum shift level and flattens out at a much later stage as compared to the other metrics. In the area of a maximum shift value from 500 to 2000 the SIFTSS is still tributing the changes in visual security with much more distinct return values than the other metrics do.

Finally, the objective image metrics' averaged return values for the Coefficient Scan Order Permutation encrypted Kodak Image Database images are shown in Figure 7. The SIFTSS values suggest an ordering with respect to visual security which corresponds to the subjective one shown in Figure 4. Contrasting to that, the return values of the PSRN, SSIM, and LFBVSM do not at all correspond to this ordering of the images. These metrics attest HP encrypted images in spatial store mode a better visual security than images which have an encrypted LP band (also in spatial store mode) which is found in the middle of the plot.

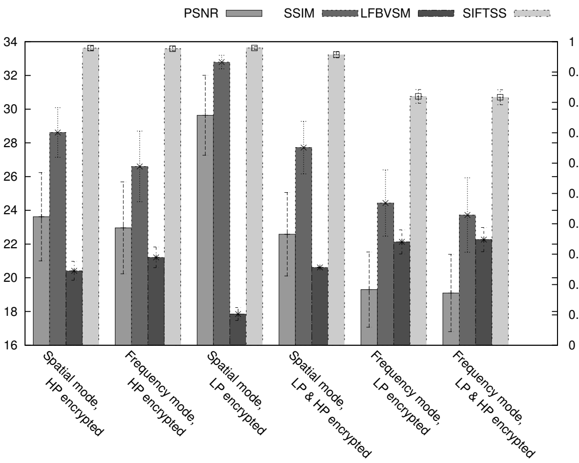

Fig. 7. Averaged return values for Coefficient Scan Order Permutation encrypted Kodak Image Database images.

Notice that the LFBVSM shows a valley or dent instead of a peak because the metric uses an inverted scale.

5 Conclusion

We have found that overall, SIFTSS is well suited to assess visual security and its ratings correspond well to subjective perception. In particular, for the three encryption techniques considered, SIFTSS clearly outperforms PSNR, SSIM, and LFBVSM in terms of correspondence to human perception and responsiveness to subtle changes in visual security.

On major drawback of SIFTSS is the computational effort required by the metric. Due to the complexity of the SIFT algorithm itself the metric is *significantly* more demanding as compared to PSNR and SSIM.

The behaviour of SIFTSS might be further improved by extracting additional characteristics from the area around each matched key-point, for example by including the mean luminance and deviation of the subregions around the key-point. These options have not been investigated so far and remain as a suggestion for future work.

References

1. Jenisch, S., Uhl, A.: A detailed evaluation of format-compliant encryption methods for JPEG XR-compressed images. EURASIP Journal on Information Security **2014**(6) (2014)
2. Au Yeung, S.K., Zhu, S., Zeng, B.: Quality assessment for a perceptual video encryption system. In: Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference on. (June 2010) 102–106

3. Khan, M.I., Jeoti, V., Malik, A.S.: On perceptual encryption: Variants of DCT block scrambling scheme for JPEG compressed images. In Kim, T.H., Pal, S.K., Grosky, W.I., Pissinou, N., Shih, T.K., Slezak, D., eds.: FGIT-SIP/MulGraB. Volume 123 of Communications in Computer and Information Science., Springer (2010) 212–223
4. Droogenbroeck, M.V., Benedett, R.: Techniques for a selective encryption of uncompressed and compressed images. In: Proceedings of ACIVS (Advanced Concepts for Intelligent Vision Systems), Ghent University, Belgium (September 2002) 90–97
5. Yeung, S.K.A., Zhu, S., Zeng, B.: Partial video encryption based on alternating transforms. IEEE Signal Processing Letters **16**(10) (October 2009) 893–896
6. Tong, L., Dai, F., Zhang, Y., Li, J.: Visual security evaluation for video encryption. In: Proceedings of the International Conference on Multimedia. MM '10, New York, NY, USA, ACM (2010) 835–838
7. Mao, Y., Wu, M.: Security evaluation for communication-friendly encryption of multimedia. In: Proceedings of the IEEE International Conference on Image Processing (ICIP'04), Singapore, IEEE Signal Processing Society (October 2004)
8. Sun, J., Xu, Z., Liu, J., Yao, Y.: An objective visual security assessment for cipher-images based on local entropy. Multimedia Tools and Applications **53**(1) (2011) 75–95
9. Yao, Y., Xu, Z., Sun, J.: Visual security assessment for cipher-images based on neighborhood similarity. Informatica **33** (2009) 69–76
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (November 2004) 91–110
11. Chen, T.Y., Huang, C.H., Chen, T.H., Liu, C.C.: Authentication of lossy compressed video data by semi-fragile watermarking. In: Proceedings of the IEEE International Conference on Image Processing, ICIP '04, Singapore, IEEE (October 2004) 2159–2162
12. Huang, L., Zhou, W., Jiang, R., Li, A.: Data quality inspection of watermarked GIS vector map. In: Proceedings of the 18th International Conference on Geoinformatics, Beijing, China (June 2010)
13. Sohn, H., De Neve, W., Ro, Y.M.: Privacy Protection in Video Surveillance Systems: Analysis of Subband-Adaptive Scrambling in JPEG XR. IEEE Transactions on Circuits and Systems for Video Technology **21**(2) (February 2011) 170–177
14. Tang, L.: Methods for encrypting and decrypting MPEG video data efficiently. In: Proceedings of the ACM Multimedia 1996, Boston, USA (November 1996) 219–229
15. Sohn, H., DeNeeve, W., Ro, Y.: Region-of-interest scrambling for scalable surveillance video using JPEG XR. In: ACM Multimedia 2009, Beijing, China (October 2009) 861–864