http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (4): 544-550, 2009
ISSN 1812-5638

A Benchmark for Perceptual Hashing based on Human Subjective Identification

[1]Hui Zhang, [2]Qiong Li, [1]Haibin Zhang and [2]Xiamu Niu
[1]Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Abstract: This study proposed a novel benchmark for evaluating the robustness and discriminability properties of perceptual hashing algorithms. Firstly, two major problems neglected by traditional benchmark are analyzed thoroughly with a concrete experiment. One problem is the inconsistence between the subjective feeling and the objective perceptual distance, the other is the partiality of the performance for different attacks. And then, in order to overcome the problems, a new benchmark for perceptual hashing based on human subjective identification is proposed and the corresponding evaluation methods are presented by illustrative experiments and examples. Present benchmark methods are fairer and more comprehensive than the traditional methods.

Key words: Perceptual hashing, benchmark, CBIR, robustness, discriminability, subjective identification

INTRODUCTION

As a new emerging technology, perceptual hashing (also called robust hashing, fingerprinting, etc.) has found extensive applications in image identification and authentication, such as CBIR, database searching, content protection and computer aided image management (Schmucker, 2006a; Swaminathan *et al*., 2006). The traditional cryptographic hash value is sensitive to the bit variation of input data. Whereas, the perceptual hash function only concerns the stable and significant perceptual content of an image and its hash vector is robust to the data format transformations and other content-preserving manipulations. As it is described by Hsu and Chun-Shein (2006), the space of queried images to a reference image can be divided into four subclasses: the original, the derived and identical versions, the derived but non-identical versions and the unrelated versions. Therefore, as shown in Fig. 1, the main goals of perceptual hashing can be depicted as:

- Make a decision near the uncertain boundary between I and {N, U}
- Measure the perceptual distance of a derived and identical version to its original

The most desired fundamental properties of perceptual hashing are the robustness and discriminability, as they are desired in the other matching problems, e.g., biometric recognition. In the context of perceptual hashing, robustness and discriminability can be defined as:

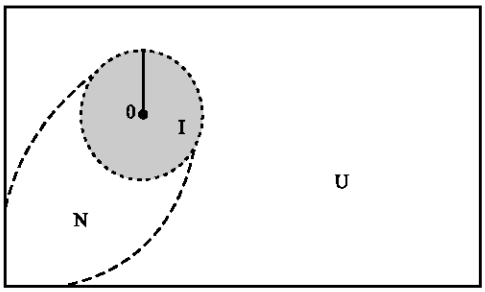

Fig. 1: Image space divided by perceptual hashing

- **Robustness:** The capability of perceptual hash to withstand the content-preserving manipulations
- **Discriminability:** The capability to distinguish the images with different perceptual contents

The dashed lines in Fig. 1 illuminate that the robustness and discriminability are conflicting in practice, for they are emphases of identifying and distinguishing cases, respectively. Namely, the identifying case emphasizes more robustness and the distinguishing case emphasizes more discriminability. Those two objects are hardly to be satisfied at the same time. Generally, the reasons for this conflict are considered to be the limitation of algorithm performance, the similarity of images, the requirements of concrete application and so on (Cano *et al*., 2002; Monga and Mhcak, 2007). However, the most important reason for this conflict is that the distortion extent of an image is not easy to be measured by an automatic quantization method.

Corresponding Author: Hui Zhang, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, China

Many well-designed perceptual hashing algorithms can be found in the published literature. Monga and Evans (2004) presented a perceptual hashing method based on geometry preserving feature points. Swaminathan *et al.* (2006) introduced a method based on the Fourier-Mellin transform and analyzed its security by the differential entropy. The algorithm of Yang *et al.* (2006) is based on the gray information in each divided block. Roy and Sun (2007) designed an algorithm based on SIFT features to detect and localize the image tampering. Venkatesan *et al.* (2000) described an algorithm utilizing the statistic of DWT coefficients and Kozat *et al.* (2004) proposed an algorithm that takes the advantage of matrix invariants of Singular Value Decomposition (SVD). For those algorithms, no matter what kind of image features are employed and what their application purposes are, the robustness and discriminability are treated as fundamental properties and evaluated by some common metrics, such as BER (Bit Error Rate), FAR (False Acceptance Rate) and FRR (False Rejection Rate), EER (Equal Error Rate), ROC (Receiver Operating Characteristic), etc. Although, these metrics and the corresponding evaluation methods have been widely used in literature, it is desired that a fair and public benchmark can be developed and used to provide more convenience comparison platform for the rapid emerging perceptual hashing algorithms. Some beginning work in this area has been done by the Watermarking Virtual Lab (WAVILA) of the ECRYPT Network of Excellence in Cryptology (Schmucker, 2006b; Schmucker and Zhang, 2006; Zhang *et al.*, 2007).

In this study, we analyzed two major problems of the traditional benchmark methods thoroughly. One is the inconsistence between the subjective feeling and the objective perceptual distance and the other is the partiality of the performance for different attacks. At the same time, we showed that the traditional benchmark methods of perceptual hashing cannot sufficiently evaluate an algorithm fairly and comprehensively. To overcome those problems, a perceptual distance measurement is given by a subjective identification experiment and a correspondingly benchmark method is proposed to evaluate the perceptual hashing algorithms.

THE ANALYSIS ON TRADITIONAL BENCHMARK

An experiment is performed to test the robustness and discriminability of two well-known perceptual hashing algorithms, denoted by Venkatesan-Wavelet (Venkatesan *et al.*, 2000) and Kozat-SVD (Kozat *et al.*, 2004) separately. An inter-class test and 6 intra-class tests are performed on the SIMPLIcity image database with 500 original images (Wang *et al.*, 2001). Each original image is attacked by 6 general attack methods with different attack strengths, they are JPEG compression, rotation, histogram equalization, Gaussian noise (by mean), Gaussian noise (by variance) and mosaic. The number of inter-class comparisons is 124,750 and the number of each intra-class comparisons for a certain attack method is 500×k, where, k, is the number of different attack strengths.

Normalized perceptual distance, optimal perceptual threshold and critical attack strength: Normally, the perceptual distance of two images given by a perceptual hashing method is measured by Hamming distance. The normalized Hamming distance can be defined as the BER:

$$BER = \frac{1}{N}\sum_{k=1}^{N} | h_1(k) - h_2(k) |$$

where, N denotes the length of the perceptual hash vectors h_1 and h_2. The BER is expected to be close to 0 when two perceptual similar images are compared and as the similarity of two images are reduced, the BER should increase accordingly.

Optimally, we assume that the perceptual hash vector generation process yields random i.i.d (independent and identically distributed) bit. When comparing two perceptually unrelated images, the random variable which denotes the number of unmatched bits, will have a binomial distribution (N, p) and the probability p of bit 0 or 1 is supposed to be 0.5. Consequently, the BER is expected to be close to 0.5 when two images with different perceptual content are compared. So, we can define the range of perceptual distance as $[d_{same}, d_{diff}]$ and for the optimal normalized Hamming distance, this range is [0, 0.5].

However, perceptual distance is not always coincident with this expected range. For some ill-designed algorithms, the distribution center of inter-class Hamming distances would deviate from 0.5. In this instance, the lower BER would not guarantee the better robustness.

A normalization method of perceptual distances is suggested as follows:

- Figure out the statistical distribution center d_{diff} of an algorithm by an inter-class test
- Determine the d_{diff} by comparing an image to itself
- Normalize a queried perceptual distance d_{same} to NPD by:

$$\frac{d - d_{same}}{d_{diff} - d_{same}} = \frac{NPD - 0}{0.5 - 0}$$

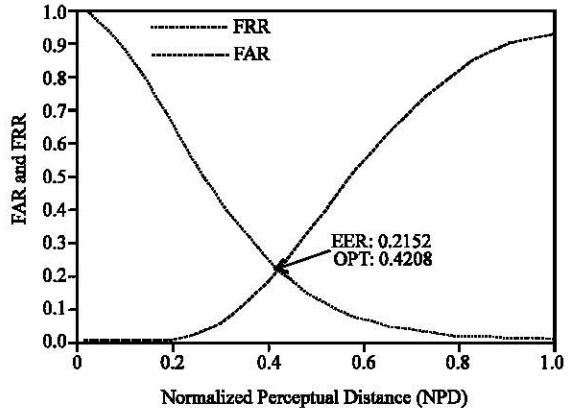

Fig. 2: The EER and OPT of Kozat-SVD under Gaussian noise (mean)

Based on the normalization method, the NPD curves can be drawn, like the traditional BER curves. In the NPD curves, NPD = 0.5 can be regard as an approximate threshold, so the lower NPD curve reflect better robustness.

From an optimal operating point f view, the EER (Equal Error Rate) is a compact term which approximately indicating the robustness and discriminability at the same time. As shown in Fig. 2, the point of EER indicated by the arrow, predicates that the sum of FAR and FRR is the minimum at this point and the corresponding perceptual distance is an optimal threshold (OPT). Therefore, we can take the OPT into account when evaluating the NPD curves. Critical Attack Strength (CAS) is defined as the attack strength corresponding to OPT in a NPD curve, which is the critical attack strength of an algorithm under a certain attack method. As shown in Fig. 3a and 4a, the algorithm with lager CAS is more robust to the tested attack.

The inconsistency between human subjective feeling and objective perceptual distance: As an instance, we compare the robustness and discriminability of Venkatesan-Wavelet and Kozat-SVD under Gaussian noise (by mean) to reveal the inconsistency in existing perceptual hashing algorithms, in term of NPD, OPT and CAS. The experimental results are shown in Fig. 3a and two examples of attacked images are listed in Fig. 3b.

Obviously, the two tested algorithms would make unreasonable decisions on the Gaussian noised baby image. As shown in Fig. 3a, the baby image noised by mean 0.3 would be regarded as non-identical by the Venkatesan-Wavelet, while the one noised by mean 0.9 would be regarded as identical by the Kozat-SVD. Comparing with the example images in Fig. 3b, we can say

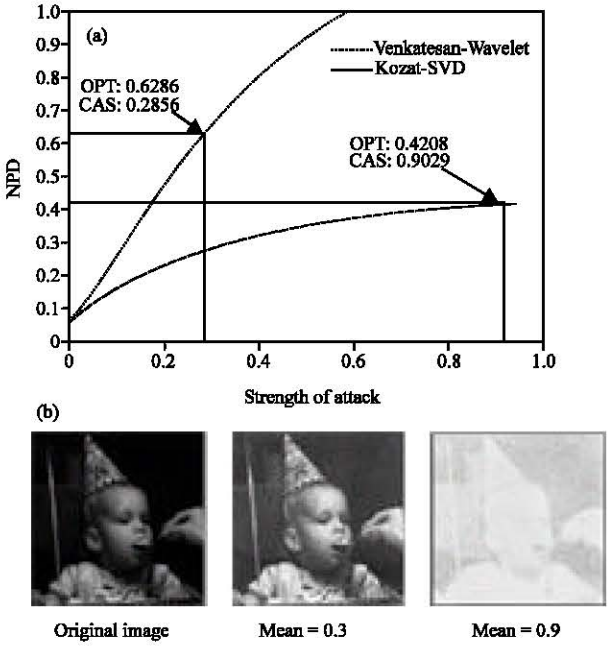

Fig. 3: Two algorithms under Gaussian noise (by mean), (a) the NPD curves of two algorithms and (b) Gaussian noised images

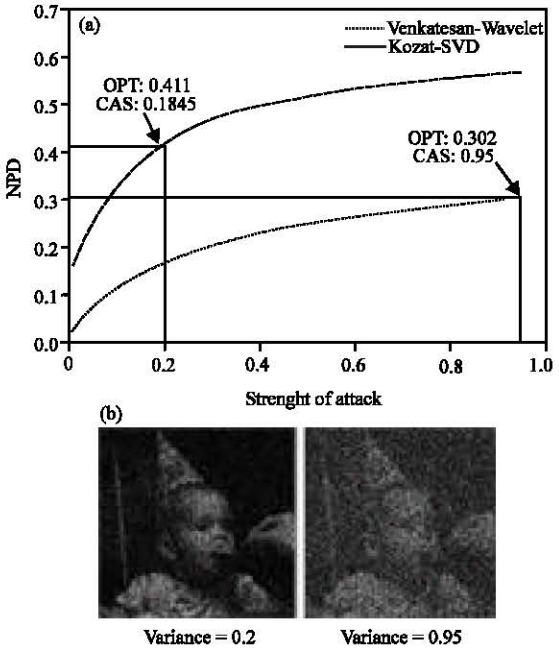

Fig. 4: Two algorithms under Gaussian noise (by variance), (a) the NPD curves of two algorithms and (b) Gaussian noised images

that the Venkatesan-Wavelet behaves too strict and the Kozat-SVD behaves too loose. However, besides those

two judgments, what judgment is reasonable and consistent with the human subjective feeling? The question can be also rephrased as: What is a reasonable tradeoff between the robustness and discriminability?

The partiality of performance: The distortions of an image in evaluation are usually simulated by some typical attack methods. For each attack method, an intra-class test is performed between the original and its variants by applying the attack with different strengths. Correspondingly, all traditional evaluation methods, such as EER, FAR, FRR and ROC curves, are applied separately on different attack methods. However, for the reasons that the image distortions are multiplex and complex in real, the separate evaluation results cannot simulate the practical case very well. For example, the performance of two algorithms under Gaussian noise (by variance) and two image examples are shown in Fig. 4a and b. Comparing Fig. 4 with Fig. 3, we can find that the performance of them present different partialities under two different noise attacks. The Venkatesan-Wavelet performs more robust under Gaussian noise (by variance). On the contrary, the Kozat-SVD performs more robust under Gaussian noise (by mean). Therefore, a uniform evaluation method based on comprehensive consideration of different attack methods is desired, The evaluation method should be able to produce more straightforward information about the performance of algorithms.

SUBJECTIVE IDENTIFICATION EXPERIMENT OF PERCEPTUAL HASHING

Because, the measure of perceptual distance should be consistent with human subjective feeling, a subjective identification experiment is proposed to get the empirical data reflecting the subject judgment about perceptual distance.

Test methodology: Similar to image quality assessment, the subjective identification experiment consults the Double-Stimulus Impairment Scale (DSIS) method (EBU method) which is described in ITU-T recommendation BT.500-11: methodology for the subjective assessment of the quality of television pictures. Since, the motivation of the experiment is to reveal a stable tolerance boundary on image quality or impairment when we identify a distorted image, a new five-grade impairment/quality scale is defined in Table 1, which enlarges the concerned scale of image distortion for the identification purpose of perceptual hashing.

Although, the image recognition process is affected by many viewing conditions, such as the display brightness and contrast, the viewing background and so on. But the most common application scenarios and the traditional evaluation conditions for perceptual hashing are very simple. Therefore, the viewing conditions in this experiment are chosen to be normal as the laboratory working conditions. The experiment was conducted in Microsoft Windows environment and with 17 inch wide screen CRT monitors at a resolution of 1440×900 pixels. The assessors are only asked to easily sit in front of a computer and rate the impairment/quality degrees of the presented attacked images according to the five-grade scale described in Table 1.

Table 1: The five grade impairment/quality scale

Rating	Impairment	Image quality
5	Imperceptible	Excellent even better quality
4	Perceptible, but not annoying	Changes slightly, but the quality is still good
3	Obviously impaired, the preserved contents are easy to be recognized	Poor visual quality, but acceptable
2	Very impaired, the preserved contents are hard to be recognized	Very bad quality, but still acceptable
1	Badly impaired, the original contents are hardly to be guessed	Unacceptable awful image quality

Fig. 5: Reference images used in subjective experiment

In this experiment, 30 reference images and 1740 destination images are rated by 68 assessors. Each assessor is asked to rate some randomly selected images within 30 min. The rating results of each test condition are recorded and averaged to generate the final Mean Opinion Score (MOS) base on the new impairment/quality scale.

Experimental image database: Thirty high quality gray-scale images are selected among the generic benchmark images. These images can reflect adequate diversity in image content, including animals, cars, airplanes, complex natural scenes and images without any specific object of interest. All of them are 512×512 pixels in size. Figure 5 shows a subset of the reference images used in this experiment.

We choose six representative image attack methods to distort these original images: JPEG compression, rotation, histogram equalization, Gaussian noise (by mean), Gaussian noise (by variance) and mosaic. The totally number of distorted images is 1740.

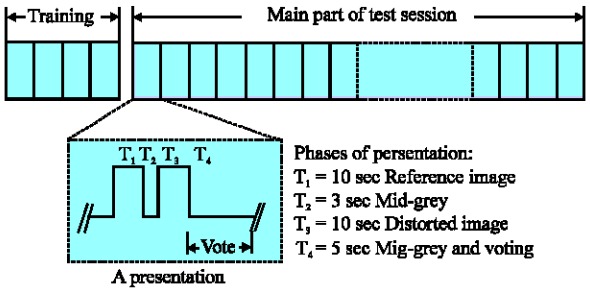

Fig. 6: Test session

The distorted images cover a broad range of impairments from high visual quality like JPEG compression with quality 95 to over distorted quality such as Gaussian noise by mean 0.95.

Test session: Each assessor is asked to participate in one test session. A test session includes a training procedure and an uncertain number of actual presentations and voting couples, as shown in Fig. 6.

Before the actual evaluation, the program will show a briefly demonstration of the test session. In the demonstration, five different reference images and their different attacked versions are presented. The participator is asked to observe the relations between the couples of images and the description of the five-grade impairment/quality scale carefully. The goal of the training phase is to make the assessors familiarize with this test environment and the rating rule. Images used in the training session will not be used in the actual test.

As shown in Fig. 6, in the main part of test session, each presentation includes four phases. The duration of each phase is definite and if the assessor cannot make his/her decision in the voting phase, the score of current presentation will be discarded. In each presentation, the reference image is randomly selected and the corresponding destination image is also selected with a random attack method and strength.

Experimental result analysis: Referring to Table 1, we make $d_{same} = 5$, $d_{diff} = 1$ and normalize the rating results to the range of NPD. Regarding the rating 2 as the threshold between identical and non-identical, we get the normalized OPT of six representative image attack methods. As a result, the corresponding subjective CAS of each attack method can be determined shown in Fig. 7a and b. As we can see, because the interval of the five grade impairment/quality scale is too large to assess the small image distortion, the value of the beginning part of the subjective test curve is relatively high.

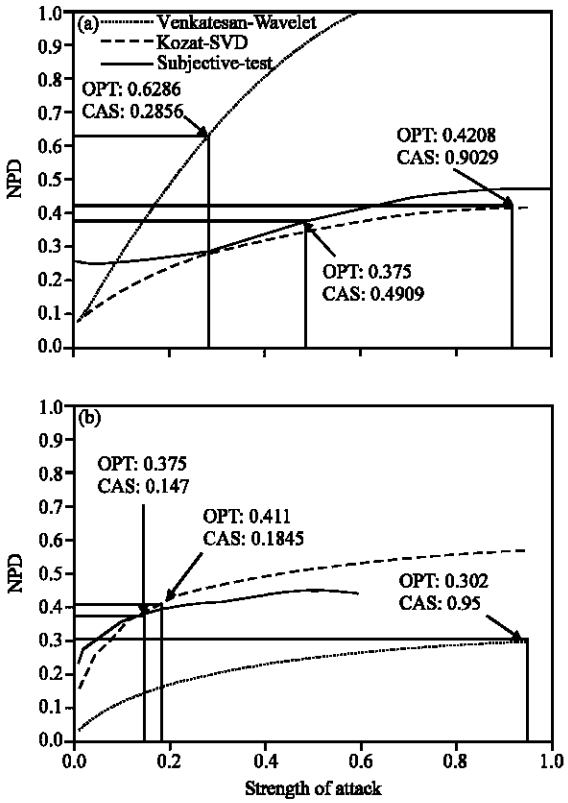

Fig. 7: Comparing the tested algorithms with the normalized subjective identification experimental results, (a) NPD curves under Gaussian noise (by mean) and (b) NPD curves under Gaussian noise (by variance)

Table 2: Comparing the tested algorithms with subjective identification experiment by OPT and CAS

Attack method	Subjective test	Venkatesan-Wavelet	Kozat-SVD
JPEG compression (quality factor)	OPT: 0.375 CAS: 2.16≈2	mNPD = 0.044 CAS: 1	mNPD = 0.116 CAS: 1
Rotation (degree)	mNPD = 0.135 CAS: 110	OPT = 0.4565 CAS: 80.4≈80	OPT = 0.513 CAS: 23.4≈23
Histogram equalization (discrete gray levels)	mNPD = 0.367 CAS: 2	OPT: 0.467 CAS: 3	OPT: 0.400 CAS: 6
Gaussian noise (variance)	OPT: 0.375 CAS: 0.147	mNPD = 0.302 CAS: 0.95	OPT: 0.411 CAS: 0.1845
Gaussian noise (mean)	OPT: 0.375 CAS: 0.491	OPT: 0.619 CAS: 0.28	OPT: 0.4208 CAS: 0.9029
Mosaic (window size)	OPT: 0.375 CAS: 6.36≈6	mNPD = 0.077 CAS: 60	mNPD = 0.270 CAS: 60

Table 2 shows the subjective CAS of each attack method. Because the content of images is reserved well after histogram equalization and rotation operations, the distorted image should be always recognized correctly by

assessors. Under such situations, we record the maximum NPD (mNPD) to reflect the worst effect of the attacks and the corresponding attack strength is also denoted with CAS.

BENCHMARK BASED ON THE SUBJECTIVE IDENTIFICATION EXPERIMENT

As it is shown in Fig. 7 and Table 2, the results of this subjective identification experiment can be used as a benchmark to evaluate the robustness and discriminability of perceptual hashing algorithms, which takes the inconsistency and the partiality into account.

Firstly, as a basic operation and facility, the NPD curve can be used to represent and compare the robustness and discriminability of the algorithms under test. The OPT is a guide to select an optimal threshold in practice and at the same time, the CAS reflect the robustness capability of an algorithm to a certain attack. The larger the CAS is, the more robust the algorithm under test is against the tested attack. In some applications, such as some occasions of computer aided image management, even if the distorted image attacked by strength near CAS is unrecognizable by human eyes, the precise distinguishing is still required.

Secondly, the proximity of the NPD curves of an algorithm and the subjective identification experimental result reflects the precision of the perceptual distance proposed by the algorithm under test. In practice, the more similar NPD curve of an algorithm is to the subjective result, the more reasonable that the perceptual distances given by the algorithm are to human subjective feeling.

Thirdly, the difference of the CAS between the algorithm and the subjective identification experimental result reflects the inconsistency of the perceptual distance and the subjective feeling. Most applications of perceptual hashing, such as CBIR, database searching and content protection, desire that the judgment of identical and non-identical can accord with the human perception as much as possible. In those cases, the small difference of CAS is desired.

Finally, if we normalize the difference of CAS under every tested attack by the following formula, a figure as Fig. 8 can be drawn to represent and compare the partialities of different algorithms in a straightforward and uniform way:

$$\frac{|CAS_{subjective} - CAS_{algorithm}|}{\text{Maximum attack strength}}$$

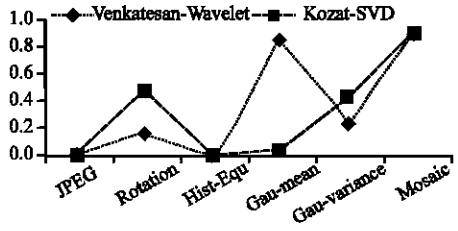

Fig. 8: Comparison of the partialities of two algorithms

DISCUSSION

With the aforementioned new benchmark methods, perceptual hashing algorithms can be evaluated and compared in more comprehensive and fairer way than with the traditional evaluation methods.

Taking the comparing of the two tested algorithms as an example, some meaningful conclusions can be drawn with Fig. 7 and 8 and Table 2 straightforwardly. Comparing the Fig. 7a and b, for the two tested Gaussian noise attacks, the perceptual distances given by the Kozat-SVD is more reasonable to human perception than that of the Venkatesan-Wavelet. But as shown in Fig. 7a, the CAS of Kozat-SVD is much larger than that of the subjective experiment. Therefore, if we use the OPT of Kozat-SVD as the threshold in practice, too many over-attacked images would be regarded as identical; while if we choose the threshold near the OPT of the subjective experiment, a relative high false rejection rate and false acceptance rate would occur. Table 2 compares the tested algorithms with subjective identification experiment in a compact way. The large CAS on mosaic of two under-test algorithms illuminates that they are both too robust to the global smooth attacks, similar to mosaic, such as blur and median filtering. On the other hand, as to the different Gaussian noise attacks, the partialities of the algorithms have been discussed with traditional evaluation methods. But the extent of the partialities can be evaluated by comparing the CAS of the algorithms with the subjective experimental results. Figure 8 represents and compares the partialities of the two algorithms on all tested 6 attack methods. For instance, a relatively lager normalized CAS of the Kozat-SVD under rotation indicates that the Kozat-SVD method is not as quite robust to the rotation attack as it is should be.

CONCLUSIONS

The present research proposed a new benchmark to provide more reasonable comparing for the application and development of perceptual hashing algorithms. Two fundamental problems concealed in traditional evaluation

methods of perceptual hashing algorithms are discussed through concrete experiments. Consequently, a subjective identification experiment is designed to get the empirical data about how people rate a distorted image. Based on the results of the subjective experiment, a new benchmark and the corresponding evaluation methods are proposed and the applications are discussed in detail with illustrative examples.

REFERENCES

Cano, P., E. Batle, T. Kalker and J. Haitsma, 2002. A review of algorithms for audio fingerprinting. Proceedings of the IEEE Workshop on Multimedia Signal Processing Lausanne, Switzerland, Dec. 9-11, IEEE, pp: 169-173.

Hsu Chao, Y. and L. Chun-Shien, 2006. Robust Image Hashing for Searching Illegal Copies. Digital Archive Task Force (DATF), Taiwan.

Kozat, S.S., R. Venkatesan and M.K. Mihcak, 2004. Robust perceptual image hashing via matrix invariants. Proceedings of the International Conference on Image Processing (ICIP). Singapore, Oct. 24-27, IEEE, pp: 3443-3446.

Monga, V. and B.L. Evans, 2004. Robust perceptual image hashing using feature points. Proceedings of the IEEE International Conference on Image Processing (ICIP). Singapore, Oct. 24-27, IEEE, pp: 677-680.

Monga, V. and M.K. Mhcak, 2007. Robust and secure image hashing via non-negative matrix factorizations. IEEE Trans. Inform. Forensics Secur., 2: 376-390.

Roy, S. and Q.B. Sun, 2007. Robust hash for detecting and localizing image tampering. Proceedings of the IEEE International Conference on Image Processing (ICIP). Hyatt Regency San Antonio, Sept. 16-19, IEEE, USA., pp: 117-120.

Schmucker, M. and H. Zhang, 2006. Benchmarking Metrics and Concepts for Perceptual Hashing. Tech. Rep., ECRYPT European Network of Excellence in Cryptology.

Schmucker, M., 2007a. Applications, application requirements and metrics. Tech. Rep., ECRYPT European Network of Excellence in Cryptology.

Schmucker, M., 2006b. Progress of forensic tracking techniques. Tech. Rep., ECRYPT European Network of Excellence in Cryptology. http://www.ecrypt.eu.org/ecrypt1/documents/D.WVL.13-1.0.pdf.

Swaminathan, A., M. Yinian and W. Min, 2006. Robust and secure image hashing. IEEE Trans. Inform. Forensics Secur., 102: 215-230.

Venkatesan, R., S.M. Koon, M.H. Jakubowski and P. Moulin, 2000. Robust image hashing. Proceedings of the International Conference on Image Processing (ICIP). Vancouver, Dec. 10-13, IEEE, Canada, BC., pp: 664-666.

Wang, J.Z., J. Li and G. Wiederhold, 2001. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Machine Intell., 23: 947-963.

Yang, B., G. Fan and X.M. Niu, 2006. Block mean value based image perceptual hashing. Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). Pasadena, California, Dec. 18-20, IEEE Computer Society, USA., pp: 167-172.

Zhang, H., M. Schmucker and X.M. Niu, 2007. The design and application of PHABS: A novel benchmark platform for perceptual hashing algorithms. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), July 2-5, IEEE, Beijing, China, pp: 887-890.