https://doi.org/10.2352/EI.2023.35.4.MWSF-375

Predicting Positions of Flipped Bits in Robust Image Hashes

Marius Leon Hammann; TU Darmstadt, Germany
Martin Steinebach, Huajian Liu, Niklas Bunzel; Fraunhofer SIT|ATHENE, Darmstadt, Germany

Abstract

Both robust and cryptographic hash methods have advantages and disadvantages. It would be ideal if robustness and cryptographic confidentiality could be combined. The problem here is that the concept of similarity of robust hashes cannot be applied to cryptographic hashes. Therefore, methods must be developed to reliably intercept the degrees of freedom of robust hashes before they are included in a cryptographic hash, but without losing their robustness. To achieve this, we need to predict the bits of a hash that are most likely to be modified, for example after a JPEG compression. We show that machine learning can be used to make a much more reliable prediction than the approaches previously discussed in the literature.

Motivation

Robust image hashing can be used for copyright protection and detection of known illegal digital images. Especially in the latter case, when private images are part of a forensic investigation, privacy is an important concern. Phones, computers and other devices are used to store many personal images whose privacy must be protected. A suspect might also not possess any illegal images at all, hence their privacy must not be compromised. While robust image hashes are highly effective in detecting known illegal images, they leak information of the original image and can thus not be seen as privacy preserving. To prevent such information leakage, Steinebach et al. propose to combine robust hashes with cryptographic hash functions into a hybrid approach [1].

A property of cryptographic hash functions is the avalanche effect, which states that a small change in the input of a cryptographic hash function results in a hash value that is drastically different from and uncorrelated with the original hash value. Therefore, distance metrics like the Hamming Distance (HD), used to match robust hashes, cannot be applied to cryptographic hashes. Consequently, an image must always produce the exact same robust hash, even when attacks like JPEG compression or rescaling are applied, or the cryptographic hashes will not match. To achieve this, Steinebach et al. propose to identify weak bits of robust hashes which can be neutralized before a cryptographic hash is applied. In this work, we improve existing approaches to predict weak bits, propose new predictions approaches based on machine learning and evaluate them.

The prediction of flipping positions is essential to combine robust and cryptographic hashes. We have discussed this topic in previous works [2] [1] . There the flipping positions are predicted by their distance to the block median value. In [3] we show that this assumption is not reliable enough for effective prediction. In our new work we use machine learning to improve the chances of correct prediction significantly. At the end, a predication will allow to combine robust and cryptographic hashing and enable superior privacy when identifying e.g. blacklisted images.

Background

Hash-based algorithms are used in various application areas, such as image search, duplicate or near duplicate detection, or image authentication. [4] [5] [6] [7] Hash functions can be divided into the two categories of cryptographic hashes and robust hashes. Cryptographic hashes are very sensitive with respect to the input data. If only 1-bit changes in the source file, when the hash is regenerated, it results in a completely new and not similar hash to the original. With a lossy compression, the original and the compressed variant would give completely different hash results. It does not matter that both images do not differ much visually. Hash-based approaches in the image context are called robust hashes or perceptual hashes. These are to be distinguished from the conventional cryptographic hash algorithms such as the MD5 hashes. Robust hashes are not very sensitive to slight modifications such as lossy compression. Even with compression, the resulting hashes would be very similar. Thus, when identifying images, the use of robust hashes is more appropriate than cryptographic hashes.

Robust Image Hashes

Robust hash functions are functions that produce a unique bit string for **perceptually** similar images. They operate based on the perceptual features of an image rather than the binary representation of the image file. Hence, they are robust to changes in single bits as long as they are not perceptually noticeable. This robustness applies to intentional and unintentional changes to an original image. These can result from malicious attempts to prevent the re-identification of an image or operations like compression and scaling, which are often applied to reduce the size of an image file during transmission. Some properties of robust hashes are:

- **Robustness**: Perceptually similar images should produce an identical or similar hash value. This includes images that have been altered to a reasonable degree, intentionally or unintentionally.
- **One-Way**: It should be impossible to reconstruct an image from a robust hash.
- **Distinction**: Perceptually distinct images should result in distinct hash values.
- **Deterministic**: The robust hashing algorithm should always produce the same hash value for a particular image.

Many different robust hash functions make use of perceptual features of images [8, 9, 10, 11]. In this work, we base our implementation on the block mean value based perceptual image hash function [12] proposed by Yang et al. in its simplest form:

- Normalize the original image into a preset size and convert it into greyscale.

- Partition the resulting image I into non-overlapping blocks $I_1, I_2, ..., I_N$ where N is the targeted length of the hash bit string.
- Permute the block sequence $\{I_1, ..., I_N\}$ based on a secret key.
- Calculate the mean pixel value of M_i of each block I_i and determine the mean value of this sequence

$$M_d = \text{median}(M_i),\ \forall i \in \{1,2,...,N\} \tag{1}$$

- Obtain binary hash value by concatenating the individual hash bits:

$$h(i) = \begin{cases} 0, & M_i < M_d,\ \forall i \in \{1,2,...,N\} \\ 1, & M_i \geq M_d,\ \forall i \in \{1,2,...,N\} \end{cases} \tag{2}$$

An example of an image and its 16x16 robust hash is shown in 1.

The robust hash applied in this work is the ForBild block hash presented by us in [13]. It is the result of an evaluation of image hashing methods [14]. Based on this hash, we have added segmentation countermeasures based on face detection [15] and watershed image segmentation [7]. Beyond the recognition of images, we also addressed the possibility of combining privacy and robust hashing in [2]. As an alternative to robust hashing, we also evaluated feature-based montage detection utilizing SIFT and SURF in [16]

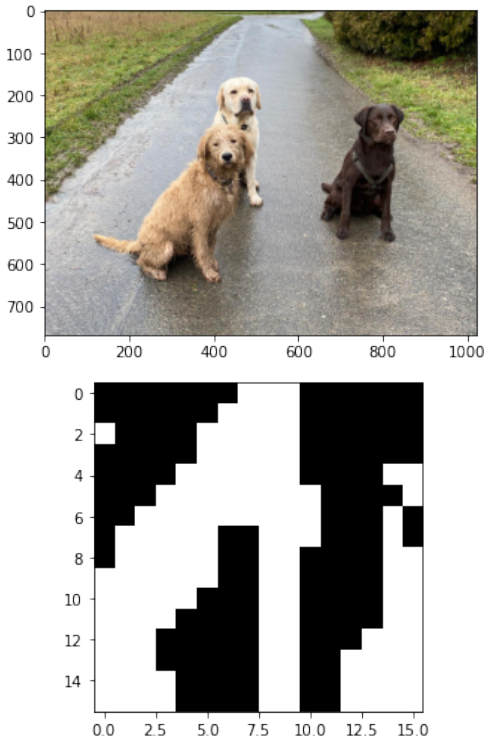

Figure 1: An image and its 16x16 robust hash

Hybrid Hash

Robust hashes can be matched not only if they are identical, they can also be similar. This similarity is measured through the hamming distance. Therefore, minor changes in the image, e.g. caused by JPEG compression or scaling, result in a robust hash that is still similar to the original robust hash. However, individual hash bits of these "attacked images" change their value depending on how much the image is attacked. This bit flipping behavior, analyzed by Steinebach et al., renders cryptographic hashes of robust hashes ineffectual because of the avalanche effect[3].
Weak bits must be predicted and neutralized before a cryptographic hash function can be applied. Steinebach et al. use the normalized distance between the pixel value of a block and the normalized median value of an image to predict such bits. The combination of robust hash, neutralization of weak bits, and a cryptographic hash function is called **Hybrid Hash**. In this work, we further analyze bit flipping during robust image hashing to improve existing heuristic predictions and propose machine learning approaches to predict weak bits.

Bit Flip Prediction by Machine Learning

One common type of problem in machine learning is classification. During classification, the model assigns the given input sample to one or more labels representing one of two or more classes. When two classes exist we speak of binary classification, or multi-class-classification when more than two classes exist. Equivalently, we speak of single-label classification when each sample is mapped to exactly one label, and of multi-label-classification when a sample can be mapped to more than one label. Predicting weak bits during robust image hashing represents a classification task. We can classify each hash block individually as either weak or stable (binary, single-label classification) or classify every hash block of a given image as weak or table (binary, multi-label-classification). In the first case of single-label-classification, the classifier receives a single block as input and returns a single label. For a block hash of size N, N classifications are required to predict weak bits of an entire image. For multi-label classification, the classifier receives an image and returns N labels, where N is the hash size of the block hash.

We present and discuss prediction approaches based on machine learning. These include a K-Nearest-Neighbor classifier, and a Deep Neural Network for multi-label classification. Additionally, we utilize a Convoluted Neural Network to predict the amount of flipped bits in an image and finally use an Convoluted Neural Network to predict the optimal tolerance for a given Quality Factor for distance-based predictions.
Images for the training and evaluation of the classifiers discussed in this section are taken from different cameras with various resolutions and show distinct motives. The same training and test set is used for every classifier. As a hash method, we use our improved block hash as introduced in [13] . The double prediction strategy discussed in this paper utilizes predictions for both the original and the suspect image.

Problem Definition

Predicting weak bits during robust image hashing represents a classification task. We can classify each hash block individually as either weak or stable (binary, single-label classification) or feed an entire image into a classifier to produce predictions for every hash bit (binary, multi-label classification). In the first case of single-label classification, the classifier receives a single block B as input and returns a single label P . For a block hash of size N, N classifications are required to predict weak bits of an entire

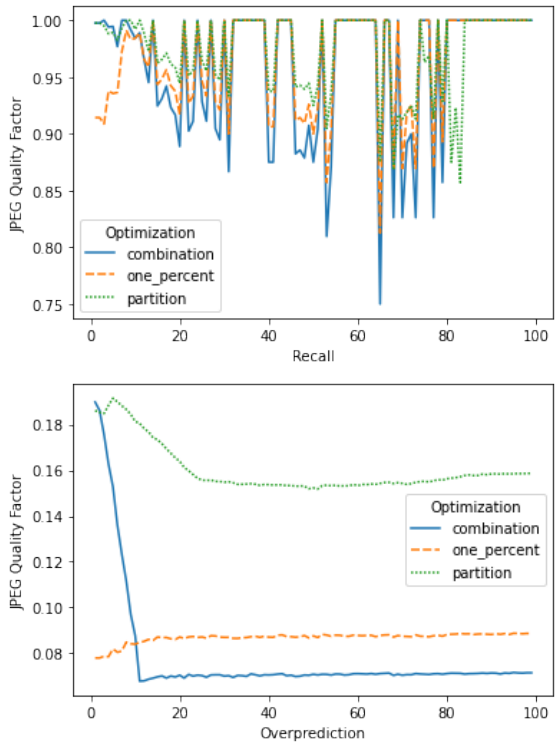

Figure 2: Recall (top) and OP (bottom) by QF for optimized kNN predictions

image. For multi-label classification, the classifier receives an image, or all blocks of an image, and returns N labels, where N is the hash size of the block hash. The classifier is called exactly once. As supervised learning is best suited for classification tasks [17], and we can generate labeled training data on demand, we use supervised learning to train our classifiers. Predicting the number of weak bits in an image or the optimal tolerance used for distance-based predictions is a regression task.

KNN

A initial analysis of the behavior of robust hash bit flip did show a dependence of the dynamic range and relative distances found in images with the likelihood of bit flipping. We, therefore, implemented a single-label classifier that classifies robust hash blocks individually based on these features. Each block can either flip (encoded as 1) or not flip (encoded as 0). The classifier predicts a class label P_N for every block B_N, for $N = 1,...,256$. We also compared the results to a classifier that considers the quality factor(qf) as a third feature. The qf is approximated using the dc coefficient of each image.
Because most blocks of a robust hash do not flip, most samples belong to class 0, which causes the dataset to be imbalanced. We fix this imbalance by re-sampling the training data s.t. both classes are represented equally.

To increase the recall of our knn classifier, we implement the following improvements:

- Combination of quality-dependent and quality-independent classifiers (“Combination”)
- Training the classifier only with low-quality images (“One Percent”)
- Replacing the continuous qf with a discrete quality value Low/Medium/High (“Partition”)

Results of this optimizations are shown in figure 2. Overall, the combined classifier showed the best mix of recall and overproduction. It is thus used as knn predictor for this work evaluated on an image level. The relative distances results in the highest recall. This shows that the differentiating between blocks with a brightness lower than the median and blocks with a brightness higher than the median improves the prediction results. On the other hand, normalized distance yields the worst results because it results in the lowest recall and a considerable over-prediction. The knn classifier shows variable recall scores likely to result in unstable hybrid hashes because of the avalanche effect. Additionally, the over-prediction is considerably high. In combination with the inconsistent recall score, the results indicate that a knn classifier is no appropriate predictor for weak bits.

Artificial Neural Network

This section introduces two Artificial Neural Network (ann) classifiers. A single-label classifier that predicts weak bits and an ann that predicts an optimal tolerance value for distance-based predictors.

Single-Label Classification

Similar to our knn prediction, we now use an ann to predict individual weak bits based on dynamic range and relative distance of each hash block. Similarly to knn predictions, a predicted class P_N is determined for every block B_N, for $N = 1,...,256$. We use a Deep Neural Network (dnn) with three hidden layers, the “adam” optimizer and the binary cross-entropy loss. The results are superior to all previous predictions in both recall (higher) and over-prediction (lower).

Again, we seek to improve recall and op by using the qf as an additional feature. The predictions using the qf achieve a nearly perfect recall score and - for qf > 10 - a significantly lower op. The higher op for low qf is acceptable because we consider such low qf to be rarely used,the nearly perfect recall at a low qf is more desirable than low op and the op is still reasonably low.

Notably, the test results show significantly lower recall scores when the amount of training epochs increases. This is likely due to overfitting on the training data. This, in turn, is likely a result of the downsampling of the training data. While more training data could be generated to allow for more training epochs, the results of our model trained with 13 epochs are already superior to all previous approaches and show little room for improvement.
One downside of this approach is that the model itself does not extract the features; thus, they must be extracted during preprocessing. Additionally, our model can only classify individual image blocks and does not consider whole images.

Tolerance Prediction

One learning from previous works [3] that the distance to the median brightness can be used to predict flipped bits with high recall. As the tolerance is used for every image and every quality factor, the op can be high, which can cause collisions when a vast dataset is used. Thus, op can be decreased when the quality factor

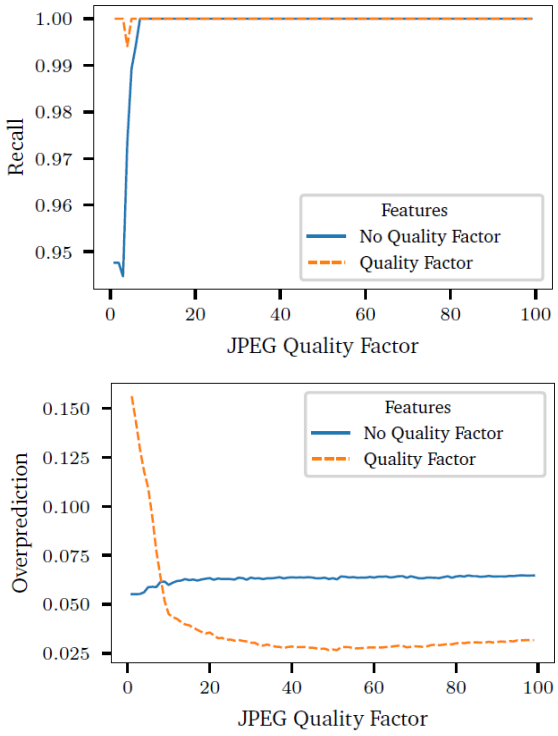

Figure 3: Recall (top) and OP (bottom) for optimized ANN predictions

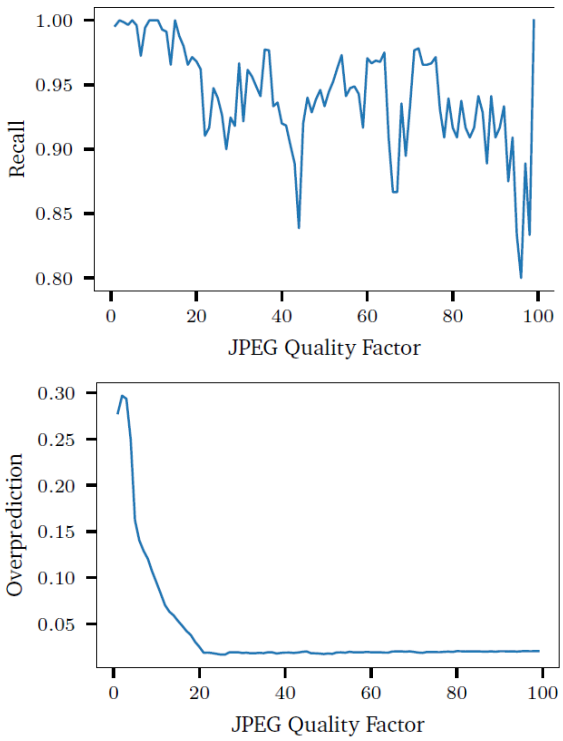

Figure 4: Recall (top) and OP (bottom) by QF for tolerance regression predictions

of an image is known. One can determine the qf based only on the first entry of the quantization table (qt). This is useful because every JPEG image requires a qt in order for it to be decodable. We use this interpolated function and train a simple neural network with only one hidden layer that takes a quantization factor as input and returns the optimal tolerance for this quantization factor. The network is trained to predict positive and negative tolerances, respectively. This approach shows a high but inconsistent recall score and the desired low op. Especially for higher qf, which are most likely to occur in a real-world scenario, the op is superior to other approaches.

Evaluation

For the evaluation of our double prediction re-identification approach we use 2000 randomly selected images of a cheerleading team from the galaxy data set[18]. The images in this data set show various amounts of humans in various poses, appearances and environments. We split these into 1000 known images and 1000 suspect images.

Results

Our evaluations show in figures 7 and 5 that our approach reaches a near perfect precision for qf independent and qf dependent re-identification against JPEG attacks. Thus, when we re-identify two images, they are very likely to be semantically the same. The recall of up to 80% shows that we do not yet detect every semantically identical image. This is especially true for low qf images. The same observations can be analogously made for the scaling attack in figures 8 and 6. The reduced recall, especially with low qf images, is caused by the fact that features of an attacked image yield different predictions from features of the original image. This shows that the features could be more robust. To further investigate this low recall we analyze the predicted bits for false negatives. Figure 9 shows the hamming distance per qf for predictions based on original images and predictions based on the respective attacked images. We considered only false negatives predicted by our dnn classifier. Attacked and original images result in similar predictions that only differ by individual bits. Because of the avalanche effect, this results in uncorrelated hybrid hashes and, ultimately, false negatives.

Summary and Conclusion

We showed that most flipped bits have low brightness distance to the median brightness and low DR. Not all image blocks with these properties result in flipped bits. Accordingly, JPEG quantization causes most flipped bits because of the quantization of the DC coefficient in the luminance channel. Correlation between JPEG QF and FR, as well as QF and the quantization factor for the DC coefficient, exist. Thus, the QT can approximate the QF, which can, in turn, approximate the FR. Our prediction algorithms use these image properties to predict weak bits in robust hashes with high recall. Our approaches can predict weak bits of an image and ultimately be used to produce privacy-preserving hybrid hashes. An open issue is that block properties like RD and DR may change during attacks. Thus, predictions for the original and attacked image do not match, resulting in a false negative. Implementing CNNs to predict weak bits from an input image could be a viable solution for this issue. In this work we considered JPEG compression and scaling as attacks. In the future additional attacks will be investigated.

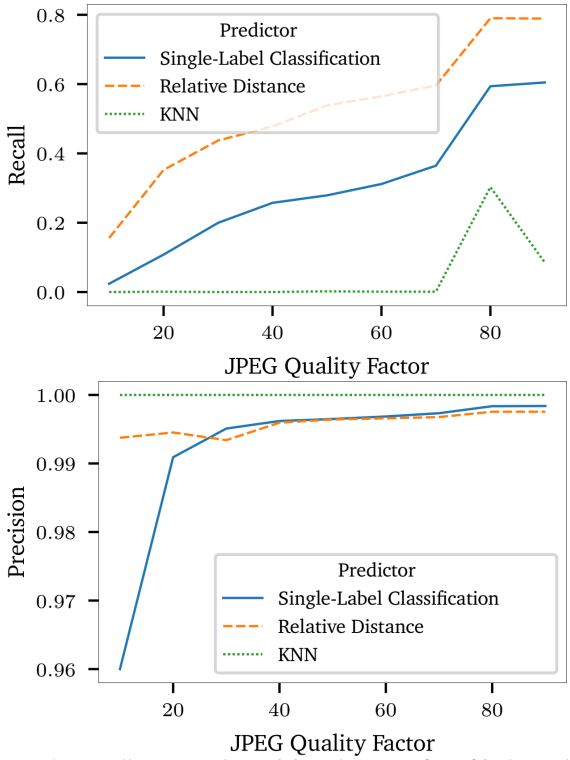

Figure 5: Recall (top) and Precision (bottom) for qf independent double prediction approach with JPEG attack.

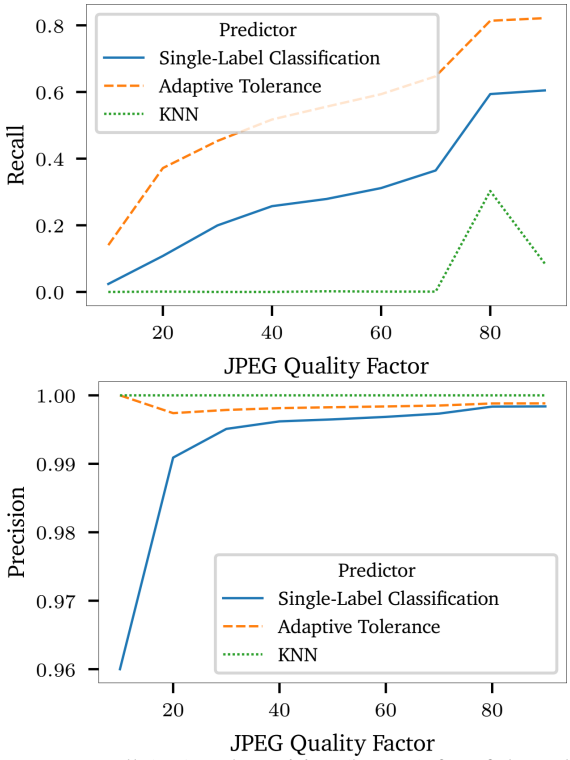

Figure 7: Recall (top) and Precision (bottom) for qf dependent double prediction approach with JPEG attack.

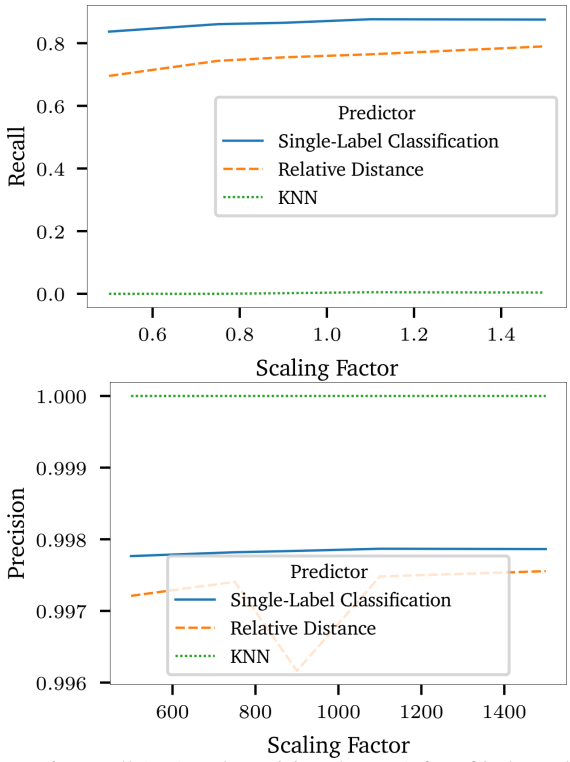

Figure 6: Recall (top) and Precision (bottom) for qf independent double prediction approach with scaling attack.

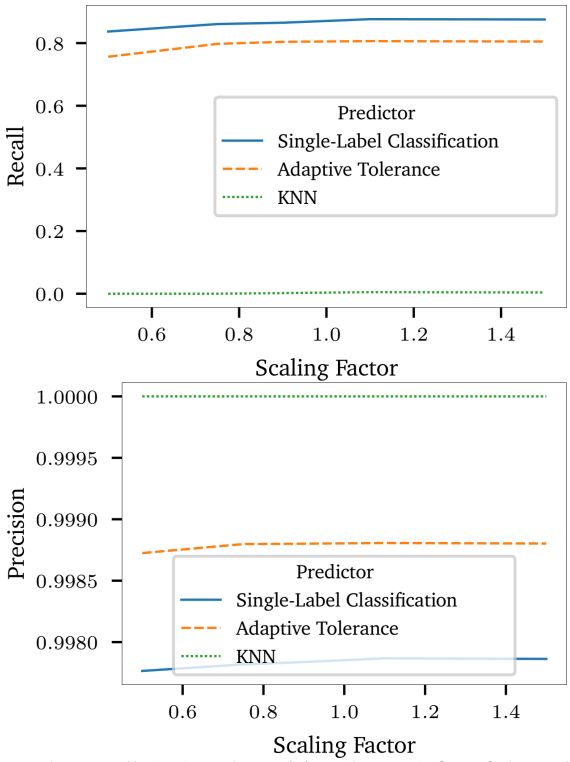

Figure 8: Recall (top) and Precision (bottom) for qf dependent double prediction approach with scaling attack.

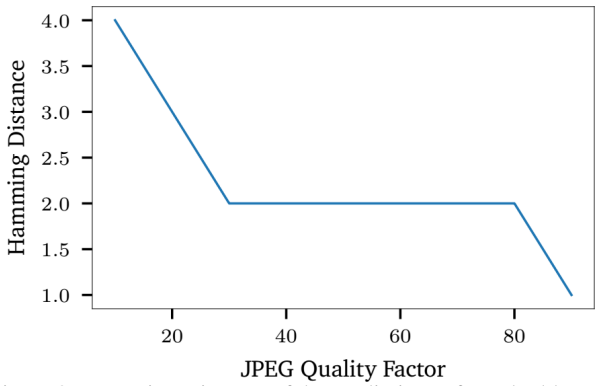

Figure 9: Hamming Distance of the predictions of our double prediction approach.

Acknowledgment

This research work has been funded by BMBF and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Martin Steinebach, Sebastian Lutz, and Huajian Liu. Privacy and robust hashes. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–8, 2019.

[2] Uwe Breidenbach, Martin Steinebach, and Huajian Liu. Privacy-enhanced robust image hashing with bloom filters. In Melanie Volkamer and Christian Wressnegger, editors, *ARES 2020: The 15th International Conference on Availability, Reliability and Security, Virtual Event, Ireland, August 25-28, 2020*, pages 56:1–56:10. ACM, 2020.

[3] Martin Steinebach. A close look at robust hash flip positions. *Electronic Imaging*, 2021(4):345–1, 2021.

[4] Andrea Drmic, Marin Silic, Goran Delac, Klemo Vladimir, and Adrian S. Kurdija. Evaluating robustness of perceptual image hashing algorithms. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 995–1000. IEEE, 2017.

[5] Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. Automatic image filtering on social networks using deep learning and perceptual hashing during crises.

[6] Ling Du, Anthony T.S. Ho, and Runmin Cong. Perceptual hashing for image authentication: A survey. *Signal Processing: Image Communication*, 81:115713, 2020.

[7] Martin Steinebach, Huajian Liu, and York Yannikos. Efficient cropping-resistant robust image hashing. In *2014 Ninth International Conference on Availability, Reliability and Security*, pages 579–585. IEEE, 2014.

[8] Zhenjun Tang, Xianquan Zhang, Xuan Dai, Jianzhong Yang, and Tianxiu Wu. Robust image hash function using local color features. *AEU - International Journal of Electronics and Communications*, 67(8):717–722, 2013.

[9] Zhenjun Tang, Fan Yang, Liyan Huang, and Xianquan Zhang. Robust image hashing with dominant dct coefficients. *Optik*, 125(18):5102–5107, 2014.

[10] Zhenjun Tang, Lv Chen, Xianquan Zhang, and Shichao Zhang. Robust image hashing with tensor decomposition. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):549–560, 2019.

[11] Rui Sun and Wenjun Zeng. Secure and robust image hashing via compressive sensing. *Multimedia Tools and Applications*, 70, 06 2012.

[12] Bian Yang, Fan Gu, and Xiamu Niu. Block mean value based image perceptual hashing. In *Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia*, IIH-MSP '06, page 167–172, USA, 2006. IEEE Computer Society.

[13] Martin Steinebach. Robust hashing for efficient forensic analysis of image sets. In Pavel Gladyshev and Marcus K. Rogers, editors, *Digital Forensics and Cyber Crime*, volume 88 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 180–187. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[14] Christoph Zauner, Martin Steinebach, and Eckehard Hermann. Rihamark: perceptual image hash benchmarking. In Nasir D. Memon, Jana Dittmann, Adnan M. Alattar, and Edward J. Delp III, editors, *Media Watermarking, Security, and Forensics III*, SPIE Proceedings, page 78800X. SPIE, 2011.

[15] Martin Steinebach, Huajian Liu, and York Yannikos. Facehash: Face detection and robust hashing. In Pavel Gladyshev, Andrew Marrington, and Ibrahim Baggili, editors, *Digital Forensics and Cyber Crime*, volume 132 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 102–115. Springer International Publishing, Cham, 2014.

[16] Martin Steinebach, Karol Gotkowski, and Hujian Liu. Fake news detection by image montage recognition. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–9, New York, NY, USA, 2019. ACM.

[17] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. A survey of machine learning for big data processing. 2016.

[18] Martin Steinebach, Huajian Liu, and York Yannikos. Forbild: Efficient robust image hashing. In *Media Watermarking, Security, and Forensics 2012*, volume 8303, page 83030O. International Society for Optics and Photonics, 2012.

Author Biography

Marius Hammann received his M.Sc. in IT Security and M.Sc. in Computer Science from TU Darmstadt in 2023. In his current position as IT Security Engineer at Aareon Group, he focuses on threat intelligence and digital forensics.

Prof. Dr. Martin Steinebach is the manager of the Media Security and IT Forensics division at Fraunhofer SIT. In 2003 he received his PhD at the Technical University of Darmstadt for this work on digital audio watermarking. In 2016 he became honorary professor at the TU Darmstadt.

Huajian Liu received his B.S. and M.S. degrees in electronic engineering from Dalian University of Technology, China, in 1999 and 2002, respectively, and his Ph.D. degree in computer science from Technical University Darmstadt, Germany, in 2008. He is currently a senior research scientist at Fraunhofer Institute for Secure Information Technology (SIT). His major research interests include information security, digital watermarking, robust hashing and digital forensics.

Niklas Bunzel received his B.Sc. and M.Sc. degrees in computer science and IT security from Technical University Darmstadt 2015 and 2020, respectively. He is currently a PhD student at the TU-Darmstadt and a research scientist at Fraunhofer Institute for Secure Information Technology (SIT) and the National Research Centre for Applied Cybersecurity - ATHENE. His major research interests include artificial intelligence, IT security and steganography.