Attacks on Digital Image Watermarks: An Experimental Study

Authors: Yuru Liu, Xuan Wu, Xiaohan Tang | Institute for Software Research | Carnegie Mellon University

Digital Image Watermarking

- Digital watermarking is the embedding of signal, secret information into the digital carrier such as image, audio and video.
- Digital image watermarking is mainly used in data authentication, copyright protection, and owner identification, etc.
- There are three categories of Algorithms: spatial domain, spectral domain, and hybrid domain

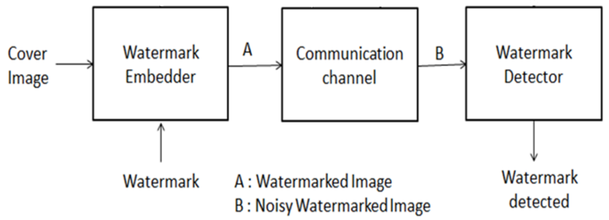


Figure 1. Digital Watermarking System

Attacks on Watermarks

- No watermarking algorithm is completely robust. Vulnerabilities vary from algorithm to algorithm.
- Malicious attacks can lead to partial or even total destruction of the embedded watermarks
- Attacks can be classified as active attacks, passive attacks, collusion attacks, and forgery attacks

Our goal

- In this project, we conduct active attacks on image watermarks from an adversary's perspective. Given an watermarked image with no other information, we try to remove the watermark.
- Through a standardized experiment, we evaluate the effectiveness of various attack methods applied at different watermarking algorithms

Experimental Design

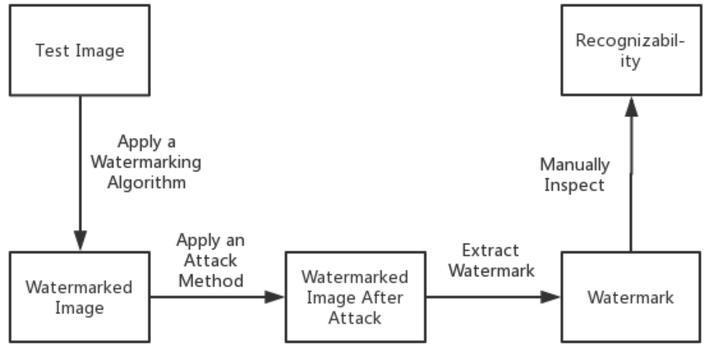


Figure 2. Flowchart of our experiment

Test Images

- 25 original image files
- 5 subject topics: scenery, portrait, objects/design, plain text, poster
- 4 formats: JPG, BMP, PNG, TIF
- 100 test images in total

Target Watermarking Algorithms

- Least Significant bit (LSB)
- Discrete Fourier Transformation Based (DFT)
- Discrete Wavelet Transformation Based (DWT)

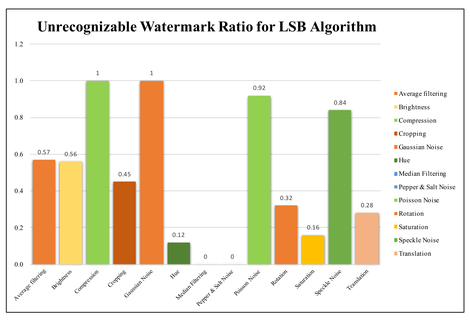
Attack Methods

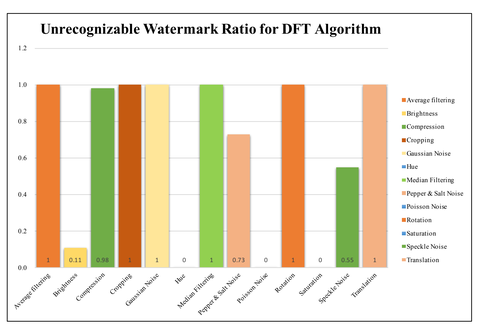
- 13 attack methods including typical geometric transformations, compression, adding noise and filtering.
- Parameters (intensity) of each method are tuned to satisfy the constraint that **PSNR** of all attacked images are above a threshold value (20 db)

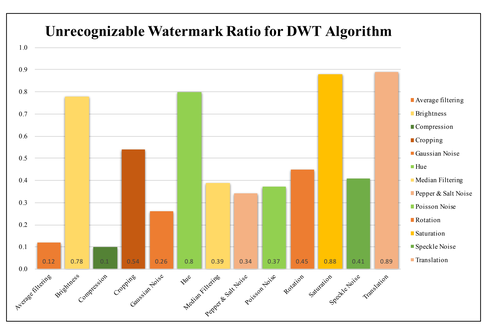
Evaluation

- For each (algorithm, attack method) pair, we inspect all the attacked watermarks with naked eye, and record the ratio of unrecognizable watermarks for comparison

Experimental Result







Conclusion and Limitations

- Watermarked Images under different themes could have divergent levels of resistance towards different attack methods. Usually a watermarked image that has high resistance towards one type of typical attack methods presents high vulnerability to another. With limited knowledge of watermarking algorithms that have been applied to the images, employing a single attack method can therefore be quite useless. Thus, combining multiple attack methods together should be the first choice when trying to remove an embedded watermark.