Chapter 1

Introduction

1.1 Basic concepts

Light: it is the electromagnetic radiation that stimulates the HVS
Luminance: it is the quantity of light perceived by the HVS
Brightness: it is the perceived luminance when objects with distinct luminance are near eache other

Hue: how much a color is ”red” (or ”blue” or ..)
Saturation: is the colorfulness of a color relative to its own brightness.

1.2 Set of attacks

- Signal processing
- Enhancement, sharpening, blurring, linear/nonlinear filtering
- Compression
- Robustness against JPEG compression is mandatory
- JPEG 2000
- Geometric manipulations
- Resizing, cropping, translation, rotation, flip

1.3 Introduction

Definition: process of secretly embedding information inside a data source

Applications

- Dissemination of digital documents
- Owner identification
- Forgery detection

- Identification of illegal copies
- Intellectual protection
- Copyright protection

Encryption does not solve the problem of unauthorized copying.
Multimedia data should be marked to allow distribution to be tracked.

Digital watermarking can provide:

- An additional layer of protection after decryption
- Data authentication and integrity

A digital watermark is an identification code bearing information (logo's, signatures) about the copyright owner, authorized consumers and so on.
It is permanently embedded into digital data for copyright protection, data authentication, integrity checking.
In most applications the water mark is not visible (perceivable) to a human observer, so that data quality is not degraded.

Copyright Protection

- A digital watermark embedded within the host signal can be retreived later to assert the owner's copyright over the marked media
- Watermark can prove ownership in court when someone has infringed on copyrights

Copy protection
The watermark represents a copy-prohibit bit and watermark detectors in the recorder determine whether the data offered to the recorder may be stored or not

Broadcasting Monitoring
Commercial advertisements and other broadcast signals could carry tracking information in the form of hidden watermarks to monitor and verify the number of times the data has been broadcast. Hence, the customer can be charged accordingly

Data Authentication
Fragile watermarks can be used to check the authenticity of the data and therefore to detect and highlight unauthorized modification to the protected data. These are weak watermarks and are designed to be destroyed in case of alteration of the marked data in an unauthorized manner

Annotation and indexing
Watermarks can also be used to annotate and index digital data. These watermarks can be used by individuals as identifiers leading to the source of the marked data or by search engines to return the marked data in relevant web searches

Medical applications
The date and the patient's details can be embedded as a watermark within medical images, which can be extracted only by authorized parties

Fingerprinting
Different watermarks in the copies of the data that are supplied to different customers (serial number)

- Copy control: to trace the source of illegal copies
- Traitor tracing: it enables the intellectual property owner to identify customers who have broken their license agreement by supplying the data to third parties

Steganography
The goal of steganography is to hide messages inside other harmless messages in a way that does not allow any enemy to even detect the presence of a second secret message.

Chapter 2

Watermarking

Def. Watermarking is a mechanism to create a communication channel that is multiplexed into original content.

Requirements:

- Imperceptibility
- Robustness: the embedded data should survive any signal processing operation the host signal goes through and preserve its fidelity
- Capacity: maximize data embedding payload
- Security - Kerckhoff's principle states that one should assume that the method used to encrypt the data is known to an unauthorized party and that the security must lie in the choice of a key

2.0.1 Blind and non-Blind

Blind techniques:
The watermark is recovered without resorting to the original non-marked image or any information derived from it

Non-blind techniques:

- the original image is needed to read the watermark
- robustness is more easily achieved

Often the application scenario does not allow the decoder to access the original non-marked image

2.0.2 Decoding process

- detectable (1-bit, 0-bit) watermark it is only possible to decide whether a given watermark is embedded in the image privateness is automatically achieved
- readable watermark (multibit watermarking) the bits hidden in the image can be read without knowing them in advance

2.0.3 Domain

Spatial domain watermarking
Watermark embedded by directly modifying the pixel values

- Simplicity
- More effective exploitation of HVS characteristics
- Syncronization problems

Transform domain watermarking
Watermark embedded in the transform domain e.g., DCT, DFT, wavelet by modifying the coefficients of global or block transformHybrid techniques

- Higher robustness
- Intrinsic resistance to some geometric manipulations
- Lack of localization
- High computational burder

Hybrid techniques
Advantages of frequency watermarking without losing spatial localization

2.1 LSB replacement

The idea of LSB watermarking is the following: we substitute the least significant k bit of the original image with the most significant k bit of the image used as a watermark.

We use two functions: *bitset(a,i,bit)*: it sets the i-th bit of a to the bit value *bitget(a,i)*: it gets the bit value of the i-th bit of a

```
function [ wmImage ] = encodeLsb(originalImage,markImage,k)

[h,w] = size(originalImage);
markImage = imresize(markImage,[h,w]);
wmImage = originalImage;

for i=1:h
    for j=1:w
        for h=1:k
            wmImage(i,j) = bitset(originalImage(i,j),h,...
            bitget(markImage(i,j),8-k+h));
        end
    end
end

end


function [rImage] = decodeLsb(wmImage,k)

[h,w] = size(wmImage);
rImage = uint8(zeros(h,w));

for i=1:h
    for j=1:w
        rImage(i,j) = uint8(bitshift(wmImage(i,j),8-k));
    end
end
```

Remark this technique is not robust as the watermark is deleted after any generic filter is applied.

2.2 Spread Spectrum Watermarking

An independent and identically distributed (i.i.d.) Gaussian random vector (the watermark) is imperceptibly inserted in a spread-spectrum-like fashion into the perceptually most significant spectral components of the data.

2.2.1 Embedding

The watermark must be composed of random numbers drawn from a Gaussian distribution $N(0, 1)$.
Embed a sequence of real values $x_1, x_2, ..., x_n$, according to $N(0, 1)$, into the n largest magnitude DCT coefficients,excluding the DC component.
We have two distinct choices:

$$Additive - SS : v_i = v_i + \alpha x_i$$

$$Multiplicative - SS : v_i = v_i(1 + \alpha x_i)$$

Remark: note that inserting αx_i means that we are modifing the frequencies and we have to be careful with respect to perceptibility. Therefore multiplicative spread spectrum is better than the additive option as we are inserting the mark proportionally to the present frequencies.

2.2.2 Detection

Compute the DCT of the watermarked (and possibly attacked) cover image I^*.
Extract the watermark X^*

$$Additive - SS : x_i^* = (v_i^* - v_i)/\alpha$$

$$Multiplicative - SS : x_i^* = (v_i^* - v_i)/\alpha v_i$$

Evaluate the similarity

$$sim(X, X^*) = \frac{X \times X^*}{\sqrt{X \times X^*}}$$

If $sim(X, X^*) > T, T$ a given threshold, the watermark exists.

The Discrete Cosine Transform (DCT)

- The DCT represents an image as a sum of sinusoids of varying magnitudes and frequencies
- The DCT has the property that, for a typical image, most of the visually significant information about the image is concentrated in just a few coefficients of the DCT
- The DCT is at the heart of the international standard lossy image compression algorithm known as JPEG

The watermark is robust to signal processing operations (such as lossy compression, filtering, digital-analog and analog-digital conversion, requantization, etc.), and common geometric transformations (such as cropping, scaling, translation, and rotation) provided that the original image is available.

Collusion

- The watermark should be robust to collusion by multiple individuals who each possess a watermarked copy of the data. The watermark should be robust to combining copies of the same data set to destroy the marks.
- It must be impossible for colluders to combine their images to generate a different valid watermarked image with the intention of framing a third party.

2.2.3 Article (Cox)

A watermark should be constructed as an independent and identically distributed (i.i.d.) Gaussian random vector that is imperceptibly inserted in a spread-spectrum-like fashion into the perceptually most significant spectral components of the data. We argue that insertion of a watermark under this regime makes the watermark robust to signal processing operations and common geometric transformations provided that the original image is available and that it can be succesfully registered against the transformed watermarked image. In these cases, the watermark detector unambiguously identifies the owner.

Introduction
Conventional cryptography provides little protection against data piracy, in which a publisher is confronted with unauthorized reproduction of information. A digital watermark is intended to complement cryptographic processes. It is a visible, or preferably invisible, identification code that is permanently embedded in the data and remains present within the data after any decryption process.

There are two parts to building a strong watermark: the watermark structure and the insertion strategy.
The stipulation that the watermark be placed in the perceptually significant components means that an attacker must target the fundamental structural components of the data, thereby heightening the chances of fidelity degradation. While this strategy may seem counterintuitive from the point of view of steganography (how can these components hide any signal?), we discovered that the significant components have a perceptual capacity that allows watermark insertion without perceptual degradation.
Further, most processing techniques applied to media data tend to leave the perceptually significant components intact.
The choice of a normal distribution gives resilient performance against collusion attacks. The Gaussian watermark also gives our scheme strong performance in the face of quantization, and may be structured to provide low false positive and false negative detection.
Nothing prevents someone from inserting another message and claiming ownership.

2.3 HVS

When hiding keep in mind the followings:

- **Keep it secret.** The hidden object is put in a place which is unknown to not authorized people. If an object location is unknown it is unlikely that it can be seen.
- **Make it small.** The hidden object is made so small that nobody is able to see it. The ability of people to perceive an object is limited by its dimension.
- **Make it similar.** The hidden object is made so similar to the surrounding environment that it is not possible to distinguish it.
- **Make it spread.** The object to be hidden is sub-divided into pieces which are spread around. In this case is the whole object that can not be perceived.

Human Visual System

- Disturbs are much less visible on highly textured regions than on uniform areas
- Contours are more sensible to noise addition that highly textured regions but less than flat areas
- Disturbs are less visible over dark and bright regions (Weber law)

2.3.1 Watermarking vs Compression

The duality between the problem of data hiding and that of compression consists in the fact that, while in compression technology the aim is to remove from the multimedia document all those data which are perceptually less important, in data hiding technology the goal is, on the contrary, to add to the multimedia document some data in such a way that they result to be perceptually unimportant.

- Compression must be very fast
- Data hiding not always have to satisfy stringent time requirements and does not require side information to be transmitted

2.3.2 Perceptual quality metrics

Subjective assessment
Visual quality is judged by human observers: observers are request to rate the quality of the watermarked image to different quality scale.
Disadvantages: it is expensive, non repeatable and it is hard to distinguish very small difference between original image and watermarked image.

Objective assessment
Error function: $e(x, y) = l(x, y) - l^{'}(x, y)$
Mean Square Error:

$$MSE = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} e(x, y)^2$$

Peak signal to noise ration

$$PSNR_{db} = 10 \log_{10} \frac{255^2}{MSE}$$

Disadvantages: it does not take into account the effect of HVS

WPSNR
Based on the fact that human eye is less sensitive to changes in textured areas than in smooth areas
WPSNR uses an additional parameter called *Noise Visibility Function* (NVF), a texture masking function, as a weight factor

$$WPSNR_{db} = 10 \log_{10} \left(\frac{255^2}{MSE * NVF^2} \right)$$

NVF uses a Gaussian model to estimate how much texture exists in any area of an image. For flat regions it assumes value 1 while 0 for edges and texture regions.

$$NVF = NORM \left(\frac{1}{1 + \delta_{block}^2} \right) \in [0, 1]$$

where $NORM$ is a normalization function and δ is the luminance variance for the block of pixel.

2.4 Theoretical models

Theoretical model 1
It is a matter of multiplexing two signals: the watermark and the host image. The problems are the following:

- non-blind: the host image has to be avaible to the detector
- blind: the host image is seen as noise (being unknown)

For blind techniques the capacity of the channel is the following

$$C = \frac{1}{2}\log\big(1 + \frac{P_w}{P_f + P_n}\big)$$

where P_w, P_n, P_f are respectively the power of watermark, noise and feature

Theoretical model 2
Defining the detector is the starting point. A watermark generator is available in order to choose the best watermark w_i depending of the feature f_i. Therefore different images will be watermarked in different ways.

Writing on dirty paper: imagine to write a message on a plain paper. Afterwards dots are added on the paper making it harder to read the message.
Suppose you know in advance where the dots will be placed on the paper. Then you will write the message where the less dots are present.
As the host image is not available at detection it is considered noise. However we know the image at the encoding side so it is possible to insert the watermark where there is less noise. By this we achieve the same capacity as in non-blind techniques (wrt to model 1) thanks to

Costa's theorem the channel capacity does not depend on f

$$C = \frac{1}{2}\log\big(1 + \frac{P_w}{P_n}\big)$$

Remark 1 $P_f \gg P_w, P_n$

Remark 2 as the watermark is inserted where there is less noise it is more sensitive to visibility.

The overall detection region is a composite set of detection regions. Note that having multiple regions brings an advantage in terms of visibility: if we had just a single region and the watermark is distant from it a lot of power would be required to make it enter the detection region and more power means more visible.

2.4.1 QIM Quantization Index Modulation

The codewords are split into groups (cosets or bins)
Every message is associated to an entire group (bin) of codewords.
The cosets must be populated enough to ensure a small distortion.

The codewords must be far enough to ensure a good degree of robustness.
Remark the trade-off is between imperceptibility and robustness.

QIM the space is quantized, for example as an n-dimensional lattice. To improve security the quantizers are randomized a bit, for example by rotating the lattice of a certain angle.

Chapter 3

Digital forensics

Why?
Modified data may influence people opinions and even alter their attitudes in response to the represented event.
It is important to be able to automatically verify the fidelity and authenticity of digital images in order to guarantee their truthfulness.
Applications of digital forensics

- Image forgery detection
- Image source identification
- Discrimination between synthetic and real images

Digital traces
Acquisition fingerprints: each component in a digital acquisition device modifies the input and leaves intrinsic fingerprints in the final output, due to the specific optical system, image sensor, camera software.

Coding fingerprints: lossy compression inevitably leaves itself characteristic footprints, which are related to the specific coding architecture

3.1 CG vs Real

3.1.1 Wavelet based approach

The decomposition of images using basis functions that are localized in spatial position, orientation, and scale (e.g., wavelet) have proven extremely useful in image compression, image coding, noise removal, and texture synthesis. One reason is that such decompositions exhibit statistical regularities that can be exploited.

The image decomposition employed here is based on separable quadrature mirror filters (QMFs). This decomposition splits the frequency space into multiple scales, and orientations (a vertical, a horizontal, and a diagonal subband). For a color (RGB) image, the decomposition is applied independently to each color channel. The resulting vertical,horizontal, and diagonal subbands for scale i are denoted $V_i^c(x,y)$,$H_i^c(x,y)$,$D_i^c(x,y)$ where $c \in \{r,g,b\}$

The first four order statistics (mean, variance, skewness, and kurtosis) of the subband coefficient histograms at each orientation, scale, and color channel are collected. These statistics form the first half of our statistical model.

While these statistics describe the basic coefficient distributions, they are unlikely to capture the strong correlations that exist across space, orientation, and scale. In order to capture some of these higher-order statistical correlations, we collect a second set of statistics that are based on the errors in a linear predictor of coefficient magnitude.

For the purpose of illustration consider first a vertical band of the green channel at scale i, $V_i^g(x,y)$. A linear predictor for the magnitude of these coefficients in a subset of all possible spatial, orientation,scale and color neighbours is given by:

$$|V_i^g(x,y)| = w_1|V_i^g(x-1,y)| + w_2|V_i^g(x+1,y)| + w_3|V_i^g(x,y-1)|+$$

$$w_4|V_i^g(x,y+1)| + w_5|V_{i+1}^g(x/2,y/2)| + w_6|D_i^g(x,y)|+$$

$$w_7|D_{i+1}^g(x/2,y/2)| + w_8|V_i^r(x,y)| + w_9|V_i^b(x,y)|$$

This linear relationship can be expressed more compactly in matrix form as:

$$\bar{v} = Q\bar{w}$$

The weights $\bar{w}$ are determined by minimizing the following quadratic error function:

$$E(\bar{w}) = [\bar{v} - Q\bar{w}]^2$$

From the measured statistics of a training set of images labeled as photorealistic or photographic, our goal is to build a classifier that can determine to which category a novel test image belongs. We employed both LDA and a non-linear SVM for the purposes of distinguishing between photorealistic and photographic images.

3.1.2 Asymmetry discrimination

To the best of our knowledge, when creating synthetic human faces, designers, in most cases, just make a half of a face and then duplicate it to create the other one. Then, they often apply post processing to achieve photorealistic results but usually not modifying the geometry of the model. Hence, if a given face present a high symmetric structure, this could be considered as a hint that it is generated via computer. On the other hand, although human faces are symmetric, there does not exist a perfectly symmetrical face.

- shape normalization
- illumination normalization
- asymmetry estimation

Asymmetry Evaluation
Let us denote the density of the image with I, and the vertically reflected of I with $I^{'}$. The edges of the densities I and $I^{'}$ are extracted and stored in I_e and $I_e^{'}$, respectively. Two measurements for the asymmetry are introduced as follows: **Density difference**

$$d(x,y) = ||I(x,y) - I^{'}(x,y)||$$

Edge orientation similarity

$$s(x,y) = \cos(\theta_{I_e(x,y),I'_e(x,y)})$$

where $\theta_{I_e(x,y),I'_e(x,y)}$ is the angle between the two edge orientations of images I_e and $I_e^{'}$, at position (x,y).

3.1.3 Expression discrimination

Generating highly realistic facial expressions is still a challenging issue, since synthetic expressions usually follow a repetitive pattern, while in natural faces the same expression is usually produced in similar but not equal ways.

Six types of facial expressions are taken into account following the six universal expressions of Ekman (happiness,sadness, disgust, surprise, anger, and fear) plus a 'neutral' one.

- From a given video sequence, frames that contain human faces are extracted
- Based on the recognition results, faces corresponding to a particular expression (e.g., happiness) are selected for the next steps. Notice that the 'neutral' expressions are not considered
- The Active Shape Model (ASM), which represents the shape of a face, is extracted from each face. In order to measure their variations, all shapes have to be comparable.
- Each extracted ASM is then normalized to a standard shape.
- Differences between normalized shapes are analysed, and based on the variation analysis results, the given sequence is confirmed to be CG or natural.

For each face, the corresponding ASM model is represented by a set of reference points, 87 in this case.
Variation analysis
The distance $d_{i,p}$ of each reference point p on a model i to the average of all points p of all models is calculated as:

$$d_{i,p} = ||(x,y)_{i,p} - (\bar{x},\bar{y})_p||$$

where $(x,y)_{i,p}$ is the position of the reference point p on the model i;
$(\bar{x},\bar{y})_p = \frac{1}{N}\sum_{i=1}^{n}(x,y)_{i,p}$, where N is the number of normalized ASM models. Depending on the facial expression ξ (among six universal expressions), a subset

S_ξ of reference points are selected for analysis.
Now let μ_p and σ_p be the mean and variance of all distances $d_{i,p}$ at reference point p over all models.
The given set of models on expression ξ is confirmed to be CG or natural by comparing the *Expression Variation Value* EVV_ξ to the threshold τ_ξ where

$$EVV_\xi = \alpha_\xi \frac{\sum_p \mu_p}{|S_\xi|\lambda_{1,\xi}} + (1 - \alpha_\xi)\frac{\max_p\{\sigma_p\}}{\lambda_{2,\xi}}$$

where α_ξ is a weighted constant and $\lambda_{1,\xi}$,$\lambda_{2,\xi}$ are the normalization values used to normalize the numerators into $[0;1]$.

3.1.4 Pulse-based discrimination

Tiny fluctuations in the appearance of a face that result from changes in blood flow. Such temporal variations are nearly invisible to the human eye, but can be revealed with "video magnification".
Because these changes result from the human pulse, they are unlikely to be found in computer generated imagery.
We use the absence or presence of this physiological signal to distinguish computer generated from human faces.

3.2 Source Identification

When shooting an image, a specific and unique fingerprint is introduced into the content, depending on the device which took it.
Image pipeline acquisition

- Lens system
  Composed of a lens and the mechanisms to control exposure, focusing, and image stabilization to collect and control the light from the scene.

- Filters
  Includes the infra-red and anti-aliasing filters to ensure maximum visible quality.

- Image sensor
  An image sensor is an array of rows and columns of photodiode elements, or pixels. When light strikes the pixel array, each pixel generates an analog signal proportional to the intensity of light, which is then converted to digital signal and processed by the DIP.

- CFA
  Since the sensor pixels are not sensitive to color, to produce a color image, a color filter array (CFA) is used in front of the sensor so that each pixel records the light intensity for a single color only.

- DIP
  The output from the sensor with a Bayer (RGB) filter (assume) is a mosaic of red, green and blue pixels of different intensities. Each pixel contains the information of only one color. The digital image processor implements interpolation (demosaicing) algorithms to recover the missing information of the other two colors for each pixel. The Bayer pattern has twice as many G filters as R or B filters because it is designed to mimic the human retina which is most sensitive to light in the green range of the spectrum. Unlike the human retina, R-G-B filters of the CFA are distributed in periodic pattern. The DIP also performs further processing such as white balancing, noise reduction, matrix manipulation, image sharpening, aperture correction, and gamma correction to produce a good quality image.

We are able to distinguish some 10 million colors from just three wavelength measurements, each made by a different type of cone: L-cone (long – orange and red), M-cone (medium - green and yellow), S-cone (short - violet and blue). Analogous to the eye's three cone types, digital cameras have three channels with peak sensitivities at different wavelengths: R, G, B. This means that each pixel is represented as three values.

Sensor imperfection
In a ideal, noise-free system, the amount of recorded light would be directly proportional to the pixel values in the final digital image. In reality, there are a variety of factors that introduce discrepancies between the amount of light that is initially recorded and the final digitized pixel values.

Matching the source by identifying and extracting systematic errors due to imaging sensor, which reveal themselves on all images acquired by the sensor in a way independent of the scene content.
Sensor's pixel defects and pattern noise (fixed pattern noise + photo response non-uniformity noise).
Detect traces of defective pixels, such as hot pixels, dead pixels, pixel traps, cluster defects.

3.2.1 Sensor noise

Each pixel in a digital camera's sensor records the amount of incident light that strikes it. Slight imperfections in manufacturing introduce small amounts of noise in the recorded image.
This noise is spatially varying and consistent over time and can therefore be used for forensics and ballistic purposes.
The image imperfections can be modeled as:

$$l(x, y) = l_0(x, y) + \gamma l_0(x, y)K(x, y) + N(x, y)$$

where l_0 is the noise free image, γ is a multiplicative constant, K is multiplicative noise PRNU (photo-response non-uniformity noise) and N is an additive noise term.

This noise arises from slight variations in the size and material properties of the sensor cells themselves. Physical inconsistencies across the sensor cells lead to differences in the efficiency with which the cells convert light into digital pixel values.

Some cells consistently under-report the amount of measured light, while others consistently over-report the amount of measured light. There variations, termed photo-response non-uniformity, lead to a stable noise pattern that is distinctive to the device.
PRNU modulates the pixel proportional to its value. It is a fixed property of the sensor and it does not vary from image to image.
The PRNU is estimated from a series of authentic images taken from the camera in question. Each image is denoised with any standard denoising filter and subtracted from the original image

$$W_k(x, y) = I_k(x, y) - I_k^{'}(x, y)$$

where $l_k^{'}$ are the denoised images. The term $W_k(x, y)$ suppress the underlying image content and make the estimation of the PRNU more reliable. The PRNU is estimated as

$$K(x, y) = \frac{\sum_k W_k(x, y) I_k(x, y)}{\sum_k I_k^2(x, y)}$$

The PRNU can then be used to determine if an image originated from a specific camera, or if a portion of an image has been altered.
For the latter application, an image in question $I(x, y)$ is denoised and subtracted from itself to yield $W(x, y)$ as described above.

The PRNU $K(x, y)$ is estimated from a set of images known to have originated from the same camera as $I(x, y)$. The correlation between the PRNU and the image being analyzed is given by

$$\rho = I(x, y)K(x, y) \otimes W(x, y)$$

where $\otimes$ denotes normalized correlation. The correlation ρ is used as a measure of authenticity and can be computed locally in order to detect localized tampering.

The best images for estimation of PRNU are those with high luminance (but not saturated) and smooth content. If the camera under investigation is in our possession, out-of-focus images of bright cloudy sky would be the best.

In practice good estimates of the fingerprint may be obtained from 20-50 natural images depending on the camera.
For color images the PRNU will be highly correlated across the RGB color channels. There is little advantage to computing the PRNU for each color channel. The RGB image can be converted to a single gray-scale image from which the PRNU is estimated.

PRNU can be used for a variety of digital forensics tasks

- Device identification
- Device linking (prove that two images were taken by the same device)
- Recovery of processing history (presence of camera fingerprint indicates that is natural and not a computer rendering - the strength or form of the fingerprint can indicate particular processing)
- Detection of digital forgeries (absence of the fingerprint in individual image regions)

3.2.2 CFA interpolation

The interpolation algorithm, used to estimate the missing color values, introduces correlation between neighbor pixels. Since CFA patterns are typically periodic these correlations will be periodic as well.
Assume that each pixel value is correlated to its neighbors with the associated weighting coefficients and each camera manufacturer uses different interpolation kernel and/or different weighting coefficients.
The crux of this approach lies on estimating these coefficients and associating them with the digital camera-model used to capture the images.
Use the outputs of Expectation/Maximization (EM) algorithm as features to detect different interpolation. Two outputs:

- The probability map. The value of each point on the probability map indicates the probability that the point is correlated with its neighbors.
- The estimate of the weighting coefficients which represent the amount of contribution from each pixel in the interpolation kernel.

- Frequency spectrum of probability maps (peaks correspond to the periodicity in probability map which reveal the CFA correlations).

- The set of weighting coefficients obtained from an image, and the peak location and magnitudes in frequency spectrum are used as features. An SVM classifier is used.

3.3 JPEG

3.3.1 JPEG compression

An image is first transformed from RGB into luminance/chrominance space (YCbCr). Each channel is then partitioned in 8×8 pixel blocks.
These values are converted from unsigned to signed integers (from $[0, 255] \rightarrow [-128, 127]$).
Each block f_c is converted to frequency space F_c using DCT.

$$F_c(\omega_k, \omega_l) = \sum_{m=0}^{7} \sum_{n=0}^{7} f_c(m, n) \cos(\omega_k m) \cos(\omega_l n)$$

where $\omega_k = 2\pi k/8$ and $\omega_l = 2\pi l/8$.
Depending on the specific frequency ω_k, ω_l and channel c each DCT coefficient F_c is quantized:

$$F_c(\omega_k, \omega_l) = round\left(\frac{F_c(\omega_k, \omega_l)}{q_c(\omega_k, \omega_l)}\right)$$

This stage is the primary source of data reduction and information loss.
The quantization is specified as a set of three 8×8 tables associated with with each frequency and image channel (YCbCr). The quantization for the luminance channel is typically less than for two chrominance channels and the quantization for the lower frequency components is typically less than for higher components.
After quantization the DCT coefficients are subject to entropy encoding (typically Huffman coding).

3.3.2 JPEG Header

The JPEG standard does not enforce any specific quantization table or Huffman code,these information are embedded in the JPEG header.
The JPEG quantization tables and Huffman codes, along with other data extracted from the JPEG header, have been found to form a distinct camera signature.
The first three components of the camera signature are:

- Dimensions [2 values], are used to distinguish between cameras with different sensor resolution
- Quantization tables [192 values]
  The set of three 8×8 quantization tables are specified as one dimensional array of 192 values
- Huffman codes [90 values]
  6 sets of 15 values corresponding to the number of words of length 1,2,..,15.
  Each channel needs two codes : one for the DC coefficient and one for the AC coefficient.

A thumbnail version of the full resolution image is often embedded in the JPEG header. It is created by cropping, filtering and down-sampling the full resolution image.

Rather than being a limitation the lack of a thumbnail is considered as a characteristic property of a camera.
Remark 284 values are extracted from the full resolution image and other 284 values are extracted from the thumbnail.
The final component of the camera signature is extracted from an image's EXIF metadata. According to the EXIF standard there are five main Image File Directories (IFDs) into which the metadata is organized:

- Primary
- Exif
- Interoperability
- Thumbnail
- GPS

It is possible to have additional IFDs. Moreover camera manufacturers customize their metadata leading to parsing errors which are considered a feature of the camera.
Remark we have 8 extracted values: 5 entries from the standard IFDs, 1 for an additional IFD, 1 for the number of entries in the additional IFD,1 for the number of parser errors.
The total is 576 values used for authentication.
Photo-editing software employs JPEG parameters that are different from the camera's; photo alteration is detected by extracting the signature from an image and comparing it to a database of known authentic camera signatures.
A pairing of camera make, model and signature is referred to as a camera configuration. All cameras with the same signature are placed into an equivalence class. An equivalence class of size n means that n cameras share the same signature. In other words an equivalence class of size greater than 1 means that there is ambiguity in identifying the camera make and model.
Remark it was found that 69% of the camera configurations are in an equivalence class of size one.

3.3.3 Double JPEG

Quantization is described as $q_a(u) = \lfloor \frac{u}{a} \rfloor$ where a is the quantization step.
De-quantization (as $q_a^{-1}(u) = au$) brings the quantized values back in the original range but it is not the inverse function of quantization.
Double quantization is quantization with step b, followed by de-quantization of step b and quantization of step a.

$$q_{ab}(u)\left\lfloor \left\lfloor \frac{u}{b} \right\rfloor \frac{b}{a} \right\rfloor$$

The double quantization of a signal introduces periodic artifacts.

3.3.4 JPEG Ghost

In JPEG quantization step each DCT coefficient c is quantized by an amount q.

$$\bar{c} = round\left(\frac{c}{q}\right)$$

Consider a set of coefficients c_1 quantized by an amlunt q_2 whihch are subsequently quantized by an amount q_2 to yield c_2.
The coefficients become increasingly more sparse as q_2 increases (change in granularity).
Remark: multiple minima may appear at values of q_2 which are integer multiples of q_1.

Instead of computing the difference between the quantized DCT coefficients we consider the difference computed directly from the pixel values (i is the color channel).

$$d(x, y, q) = \frac{1}{3}\sum_{i=1}^{3}[f(x, y, i) - f_q(x, y, i)]$$

The difference will change significantly from highly textured regions to flat regions. The solution is to average over $b \times b$ pixel regions.

3.4 Geometric based forensics

3.4.1 Detectiong composites of people

The compositing of two or more people into a single image is a common form of manipulation.
The following method estimates a camera's principal point from the image of a person's eyes. Inconsistencies in the principal point are then used as evidence of tampering.The key fact is that translation in the image plane is equivalent to a shift of the principal point.

Method: In general, the mapping between points in 3-D world coordinates to 2-D image coordinates is described by the projective imaging equation: $\mathbf{x} = P\mathbf{X}$ where the matrix P is a 3x4 projective transform, the vector X represents a world point in homogeneous coordinates, and the vector x represents an image.
If all the world points $\mathbf{X}$ are coplanar, then the world coordinate system can be defined such that the points lie on the $Z = 0$ plane. In this case, the projective transformation P reduces to a 3 x 3 planar projective transform H, also known as a **homography**: $\mathbf{x} = H\mathbf{X}$.

Homography estimation: the homography H between points on a world plane and its projection on the image plane can be estimated if there is known geometry in the world: parallel lines, orthogonal lines, regular polygons, or circles.
We describe how to estimate a camera's principal point from the image of a pair of eyes. A simple 3-D model for an eye consists of two spheres. The larger sphere, with radius $r_1 = 11.5mm$, represents the sclera and the smaller sphere,

with radius $r_2 = 7.8mm$, represents the cornea. The *limbus*, the boundary between the iris and the sclera, is defined by the intersection of two spheres, a circle with radius $p = 5.8mm$.

Camera calibration: once estimated, the homography H can be decomposed in terms of its intrinsic and extrinsic camera parameters.
The intrinsic parameters consist of the focal length f , principal point (c_1, c_2), skew σ, and aspect ratio α. The extrinsic parameters consist of a rotation matrix R and translation vector t that define the transformation between the world and camera coordinate systems. Since the world points lie on a single plane, H can be decomposed in terms of the intrinsic and extrinsic parameters as: $H = \lambda K(r_1 r_2 t)$ where λ is a scale factor and the 3×3 instrinsic matrix K is:

$$\begin{pmatrix} \alpha f & \sigma & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{pmatrix}$$

For simplicity we will assume the skew $\sigma = 0$ and the aspect ration $\alpha = 1$. under these assumptions matrix K is

$$\begin{pmatrix} f & 0 & c_1 \\ 0 & f & c_2 \\ 0 & 0 & 1 \end{pmatrix}$$

With only two constraints, it is possible to estimate the principal point (c_1, c_2) or the focal length f , but not both.

Translation: In homogeneous coordinates, translations are rep- resented by multiplication with a translation matrix T:

$$\begin{pmatrix} 1 & 0 & d_1 \\ 0 & 1 & d_2 \\ 0 & 0 & 1 \end{pmatrix}$$

The mapping from world $\mathbf{X}$ to (translated) image coordinates y is:
$y = TH\mathbf{X} = \lambda TK(r_1 r_2 t)\mathbf{X} = \lambda \bar{K}(r_1 r_2 t)\mathbf{X}$ where:

$$\bar{K} = \begin{pmatrix} f & 0 & c_1 + d_1 \\ 0 & f & c_2 + d_2 \\ 0 & 0 & 1 \end{pmatrix}$$

3.4.2 Detecting photo manipulation on signs and billboard

The manipulation of text on a sign or billboard is relatively easy to do in a way that is perceptually convincing. When text is on a planar surface and imaged under perspective projection, the text undergoes a specific distortion. When text is manipulated, it is unlikely to precisely satisfy this geometric mapping. We describe a technique for detecting if text in an image obeys the expected perspective projection, deviations from which are used as evidence of tampering.

When inserting text into an image it is likely that the precise rules of perspective projection will be violated, and that these violations will not be perceptually

obvious.
This method explicitly identifies the projection of text on a planar surface and detects deviations from this model. We can distinguish two cases: the font style is know or not.

Planar homography: we consider a special case of this geometric transform where all of the world points X lie on a single plane and P reduces to a 3x3 planar projective transform H, known as a homography: $\bar{x} = H\bar{X}$. this equation can be rewritten as $A\bar{h} = \bar{0}$ where A is a 3×9 matrix. Note that the rows of A are not linearly independent and this system provides two constraints in eight unknowns. In order to solve for $\bar{h}$, we require four or more points with known image, $\bar{x}$, and (planar) world, $\bar{X}$, coordinates.
We employ the SIFT operator to extract the coordinates of distinctive image keypoint positions. These keypoints are invariant to certain amounts of image scale, rotation, affine distortion, noise, and illumination differences.

the inverse homography is applied to the keypoints in image coordinates yielding rectified world coordinates: $\bar{X_r} = H^{-1}\bar{x}$.
It is unlikely in an inauthentic image to have the image coordinates precisely satisfy the proper planar perspective distortion. In this case, the rectified image $I_r(x, y)$ is unlikely to match the world image $I_w(x, y)$. On the other hand, in the case of an authentic image, the rectified image should be a good approximation of the world image. As such, we use the root mean square (RMS) error between the world and rectified image as a measure of authenticity:

$$e = \frac{1}{\sqrt{n_x n_y}} ||I_w - I_r||$$

where n_x, n_y are the image dimensions.