Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Interpol review of imaging and video 2016–2019

Zeno Geradts PhD Senior Forensic Scientist *, Nienke Filius, MFS, Arnout Ruifrok PhD Senior Forensic Scientist

Netherlands Forensic Institute, Laan van Ypenburg 6, 2497, GB Den Haag, Netherlands

ARTICLE INFO

Article history:
Received 6 January 2020
Accepted 23 January 2020
Available online 30 May 2020

Keywords:
Imaging
Video
Image analysis
Interpol

ABSTRACT

This review paper covers the forensic-relevant literature in imaging and video analysis from 2016 to 2019 as a part of the 19th Interpol International Forensic Science Managers Symposium. The review papers are also available at the Interpol website at: https://www.interpol.int/content/download/14458/file/Interpol%20Review%20Papers%202019.pdf.

1. Introduction

In this review, the most important developments are presented for the following general fields of expertise:: (1) Detection of image manipulation, (2) biometric comparison (face, gait and other biometrics) (3) Camera source identification.

2. Working groups and organizations

The development of forensic image analysis has several international working groups:

- **OSAC Digital/Multimedia Scientific Area Committee**: previously the SWGIT an American group that has produced a lot of guidelines and best practice manuals. http://www.swigit.org The group has terminated operations, since the OSACs are formed http://www.nist.gov/forensics/osac.cfm.
- **ENFSI DIWG**: The ENFSI Digital Imaging Working Group that is focused on methods, techniques, education and training.http://www.enfsi.org
- **LEVA**: an American group focused on video processing and training: http://www.leva.org
- **AGIB**, A working group in Germany that is focused on facial image comparison: http://www.foto-identifikation.de/.
- **FISWG**, An American group since 2009 that is focused on facial image comparison: http://www.fiswg.org
- **OSAC Facial Identification Subcommittee**, An American group part of the Organisation of Scientific Advice Committees, with focus on standards and guidelines related to the image-based comparisons of human facial features: http://www.nist.gov/forensics/osac/sub-face.cfm

2.1. American academy of Forensic Science [36]

Within the American Academy of Forensic Science the Digital and Multimedia Sciences Section works in this field.

Since 2003 each year a workshop was organized on Forensic Image and Video processing with handouts on the methods for face comparison, video restoration, 3D reconstruction, length measurement, photogrammetry and image processing. Also each year a scientific session was organized on this field. More information is available on: http://www.aafs.org.

2.2. ENFSI forensic IT working group

The forensic IT working group of ENFSI [37,38] deals with digital evidence as such. There exist some overlap with the Digital Imaging working group, and for that reason joint events are organized.

Since most CCTV-systems are digital nowadays, often the question of handling the CCTV system itself is a question of digital evidence. Hard drives and other digital media should be handled in a secure way with proper forensic imaging software. The working

* Corresponding author.
E-mail addresses: z.geradts@nfi.nl (Z. Geradts), a.ruifrok@nfi.nl (A. Ruifrok).

https://doi.org/10.1016/j.fsisyn.2020.01.017
2589-871X/

group organizes training conferences each year. More information is available from http://www.enfsi.eu/.

2.3. *Outline of this work*

Since in the field of forensic image and video investigation there are many new developments and in the literature over 3000 references could be found in the last three years, in this review we focus only on specific areas. The area of image manipulation detection as well as the deepfakes which have given much attention the last years as well as the developments in facial and biometric comparison. Most of the developments are related to deep learning algorithms, so for this reason Nienke Filius worked on a review of the literature in the last three years which is included in chapter 4. Chapter 5 handles images and video in biometrics, whereas chapter 6 discusses camera identification.

2.4. *Detection of image manipulation*

In this chapter we go in depth on digital image manipulation and deep learning. Since deep learning is a major development in the field, this is a starting point in new literature.

Digital images and videos provide us with an effective and natural medium for communication, due to its immediateness and easy way to understand the content. As such digital images have taken on an important role in a broad range of applications. They are widely used in news reports, as evidence in legal proceedings and criminal investigations, for medical imaging, and for signal intelligence in military and governmental scenario's [1,2]. However, with the rapid spread of low-cost and easy to use devices for the capturing of visual data, almost everybody has the accessibility of recording, storing and distributing large amounts of data. At the same time, the availability of low-cost, user friendly image editing software make it extremely easy to create, alter and modify the information represented by an image without leaving any traces visible to the human eye (see Fig. 1) [1,3,4]. The art of manipulating and counterfeiting visual content is no longer restricted to experts only.

A digital image may go during its lifetime, from capturing to presentation, through a number of processing steps intended to, for example, to enhance the quality of the image. However, these steps could also include actions with the intend to tamper with the content or to create new content by combining pre-existing material. Manipulated images are appearing with increasing frequency and can have important consequences for governmental, commercial, and social institutions who rely on digital images for information [4,5]. If for example a manipulated photo is used as evidence in a legal proceeding it could lead to a misjudgement of justice. To regain trust in the authenticity and truthfulness of digital imaging researchers have developed a wide range of techniques for the detection of image manipulation and for the reconstruction of an image processing history [2,4].

There are two main questions that arise when we want to verify the history and authenticity of a digital image: ‘Was the image captured by the device it is claimed to be captured with?’ and ‘Does the image still depict its original content?’ [3]. The first question is of interest when the device suspect of capturing the image represents the evidence itself. The second question is of a more general interest and the answer to that questions can be relatively simple when the original image is known. However, in reality almost no information of the original image can be assumed to be known in advance, through research the authenticity of the image has to be verified in a ‘blind’ way [3,4]. To solve the issue put forward by the second question is the main goal in image manipulation detection research.

Image manipulation detection methods can be categorised into two main categories: (1) active and (2) passive or blind. Active manipulation detection techniques, such as digital watermarking, make use of an authentication code that is embedded into the image's content before the image is sent. The authenticity of the image is than verified by comparing the authentication code to the original code [1].

Passive manipulation detection techniques make use of the actual digital image itself to assess its credibility. This technique is based on the assumption that although the digital manipulation may not leave any traces visible to the human eye, the manipulation probably does disturb the underlying statistical properties or consistencies. This will introduce artefacts that result in various forms of irregularities. These irregularities can subsequently be used to detect the manipulation operations applied [1].

2.4.1. *Copy-move*

Copy-moves is one of the most common image manipulation technique used due to is simplicity and effectiveness. In copy-move part of the original image is copied (cloned), moved to the desired location within the original image, and pasted. It is mostly used to hide certain details or to duplicate certain aspects of an image. Textured regions are ideal for copy-move forgery. They have similar colour and noise variation properties to that of the original image which are unperceivable to the human eye looking for inconsistencies in the image statistical properties. Blurring is usually applied along the boundary of the modified region to reduce the

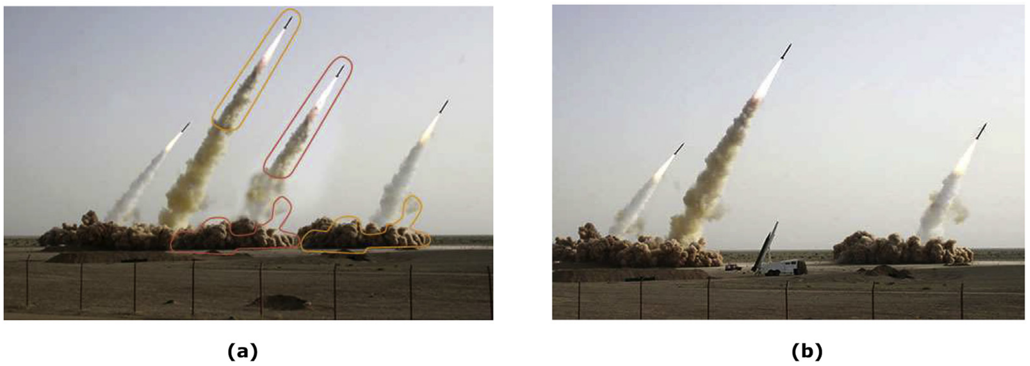


Fig. 1. Example of image manipulation that appeared in press in July 2008. (a) The forged image displaying four missiles. Only three of them are real, two different sections (encircled in red and orange, respectively) are replicates of other image sections (b)The original image showing only three missiles [6]. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

effect of irregularities between the original and pasted region [1].

2.4.2. Splicing

Splicing, or cut-and-paste, is used to modify the composition of an image by using fragments of one or more different images and paste them into another image. Geometric transforms (e.g., scaling or rotating) are often applied to make sure the pasted fragment compliments the perspective and scale of the original image [3].

2.4.3. JPEG compression properties

Identifying whether or not an image has been previously JPEG compressed plays an important role in image manipulation detection. After editing, the image is often saved in JPEG format and as such re-compressed. This second JPEG compression will introduce a deviating fingerprint when compared to single compression [4].

2.4.4. Median filtering

Median filtering is mostly used as an anti-forensic technique. Anti-forensic techniques are techniques applied by the forger to hide or remove traces left by certain image manipulation operations [4,7]. A median filter can smooth artefacts of JPEG-compression and geometric transforms or remove impulsive noise. Median filtering is a filter that operates by using a sliding window, also known as a kernel, that moves over the image while keeping the median pixel value within the window's dimension [2].

2.4.5. Local noise

Original images have an amount of noise that is uniformly distributed across the entire image. A common anti-forensic technique is to add localised random noise to the image regions that are tampered with to conceal traces of manipulation. The detection of inconsistent local noise levels over the image can be used to detect image manipulation [1].

Other techniques used to hide the manipulation operations to the human eye are enhancements such as sharpening, contrast adjustment and colour modification [1,4].

Most research in image manipulation detection was focused on detection of traces left by one specific editing operation (e.g. copy-move, JPEG compression, resampling, contrast enhancement), then they developed algorithms to detect the statistical characteristics that could reveal these traces [8]. The development of these targeted manipulation detection techniques have led to many important advances in image manipulation detection. However this approach has an important drawback: forgers have many manipulation operations at their disposal. To determine if and how an image was manipulated the forensic investigator has to apply numerous forensic tests. The need to run multiple forensic detection tests on an image to detect image manipulation confronts the investigator with new problems. For instance, how to control for the overall false alarm rate between multiple tests or how to handle conflicting results. And as new image manipulation operations are unfold, the traces left by these new operations need to be identified and an associated detection algorithms needs to be developed [9], which is both difficult and time consuming.

Therefore, there is a growing interest in the evolvement of universal forensic detection algorithms, designed to detect many, if not all, manipulation operations. The introduction of deep learning and convolutions neural networks (CNNs) has fuelled these developments. CNNs have the ability to adaptively learn classification features from large sets of data, instead of relying on humanly selected features [9,10]. The objective of this report is to provide an overview of the developments in the last three years in convolutional neural networks for universal image manipulation detection.

The report is build up as follows. The next chapter gives a brief overview of convolutional neural networks. The third chapter gives a summary of the recent developments in universal image manipulation detection using CNN's. And the fourth chapter discusses the benefits and drawbacks of the different CNN architectures with recommendations for future research.

3. Convolutional neural networks

A convolutional neural network (CNN) is very similar to a regular multi-layer neural network with the exception that it makes the explicit assumption that the input is an image. CNN's take advantage of this assumption by constraining the architecture in a more sensible way. Unlike regular neural networks with neurons in the convolutional layer arranged in one dimension, the convolutional layers of a CNN have neurons arranged in three dimensions: width, height and depth as can be seen in Fig. 2. These dimensions refer to the dimensions of the image. Every layer transforms the 3D input volume of the image to a 3D output volume called the feature map [11].

Although the particular design of CNN's may differ, they are built using a common set of basic elements. As a result the CNN's share a similar overall architecture [9]. A convolutional neural network is build of three main layer categories: convolutional layer(s), pooling layer(s) and fully connected layer(s) stacked together to form al full convolutional neural network [11].

The convolutional layer is the core building block of a convolutional neural network. Every convolutional layer consists of one or more learnable convolutional filters (i.e. a filter with learnable weights and biases). Every filter (or kernel) is small in the spatial dimension (width and height), but extends through the full depth of the input image [11]. For example, a typical filter of the first convolutional layer might have size 5 × 5 × 3 (i.e. 5 pixels width and height, and an image depth of 3 corresponding to the three RGB colour channels). Each filter is slid (or more precisely convolved) across the width and height of the input image. As the filter is slid over the input image a 2-dimensional activation map is produced that gives the response of that filter at every spatial position [11]. The windows of the filter positions can overlap, the overlapping distance is called the stride [12]. In each convolutional layer we have an of set filters and each of them produces a 2D activation map. The activation map of each filter is stacked along the depth dimension to produce the output volume, known as feature maps. These filters serve as a set of feature extractors and the convolutional layers are trained to automatically learn filters that activate when they see some type of feature [9,11]. The activation maps are often followed by activation functions, such as rectified linear unit (ReLU), exponential linear unit (ELU), Parametric ReLU (PReLu) or hyperbolic tangent (Tanh). The activation function introduce non-linearity [12].

The pooling layers function is to progressively reduce the spatial size of the feature map to reduce the amount of parameters and computational costs of training the network, and thus to control overfitting [9,11]. The pooling layer performs a down sampling operation along the spatial dimensions (width, height) of the feature maps. It operates by sliding a filter over the feature map with overlapping windows, only maintaining a single value per window for every depth slice. Resulting in a volume of smaller size, but with the depth dimension unchanged [9,11]. There exist many types of pooling operations. Two of the most popular are average pooling and max pooling. With average pooling the mean value of each window is retained and with maximum pooling the maximum value of each window is retained [9,12]. Most CNN's are built using a combination of convolutional layers and pooling layers stacked on top of one another. “This enables the CNN to learn a set of low-level features in early layers, then hierarchically group them

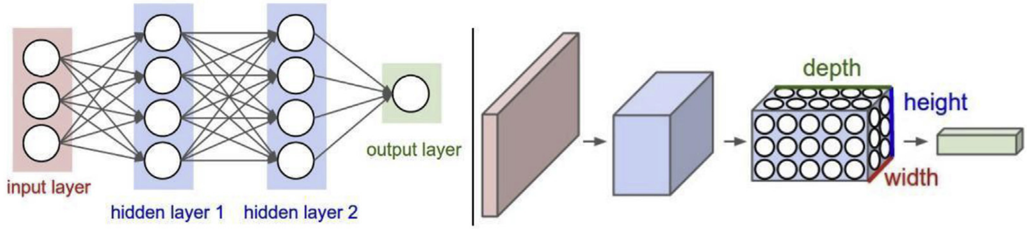


Fig. 2. Left: A regular 3-layer (2 hidden and 1 output) neural network with one dimensional layers. Right: a convolutional neural network with the neurons arranged in three dimensions. Every layer transforms the 3D input volume to a 3D output volume of neuron activation's. The red input layer holds the image, with its width and height equal to the spatial dimensions of the image, and a depth of 3 (the Colour channels Red, Green, Blue) [11]. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

into high-level features in later layers" [9]. The output is a final set of feature maps that is passed on to the fully connected layers to perform the ultimate classification.

Equal to regular neural networks, each neuron in the fully connected layer is connected to all neurons in the preceding layer [9,11]. Multiple fully connected layers can be put one after another to create deep architectures. The ultimate fully connected layer, (or output layer) has one neuron coinciding with each possible classification. The output of the ultimate fully connected layer is usually passed on to a softmax function that maps the classifications to a set of probability values such that the total sum of the output is equal to one [12]. It tells you the probability that any of the classifications is true.

At the start of the training process the filters coefficients are initially seeded with random values. During CNN training the coefficients of the convolutional filters in the network are automatically learned using an iterative algorithm that alternates between feed-forward and back-propagation runs of the data. The aim of the algorithm is to minimise the average loss between the true classification and the network output [9]. When training the CNN is finished, the CNN is tested by feeding the CNN with a test set and analysing the results by calculating the accuracy. The accuracy is the proportion of correct classifications among the total cases tested.

3.1. Recent developments

Convolutional neural networks (CNN's) have fuelled substantial advances in image recognition due to their capability to adaptively learn strong classification features for object recognition. However, in their existing form CNN's are not well suited for image manipulation detection [9]. The main difference between image recognition and image manipulation detection is the signal strength. Image manipulation detection, in contrast to image recognition, has to cope with very small differences between the manipulated image and the original image [13].

This issue was recognised by Chen et al. [14], one of the first to use CNN's for image manipulation detection. Chen et al. [14] proposed a CNN model for the detection of different values of median filtering (3 × 3 and 5 × 5). In their initial experiments conventional CNN models were directly employed as median filtering forensic models (i.e. the raw image pixels were used as input to the CNN's). These models did not perform well, suggesting that existing CNN models have difficulty to capture the important statistical forensic properties [14]. Thus, when using the standard architecture the convolutional layers tend to extract features that capture an image's content, instead of identifying traces left by editing and manipulation [10]. This led researches to investigate CNN architectures and adapt them to make them suitable for image manipulation detection.

4. CNN architecture

4.1. Preprocessing layer

The need in image manipulation detection to suppress an images content and capture the pixel value dependencies induced by manipulation operations led Chen et al. [14] to propose a modification to the conventional CNN model: adding a filtering layer. This filtering layer outputs the median filtering residual (MFR) of an image, thereby suppressing the interference caused by image edges and textures. The output MFR is fed into a traditional convolutional neural network consisting of 5 convolutional layers with ReLU activation function, followed by max pooling layers after the first, second and fifth convolutional layer and three fully connected layers with softmax activation for classification. The input to their model were gray-scale images sized 64 × 64 and 32 × 32.

The proposed model was trained and tested to detect median filtering with a binary classification approach (original/manipulated), instead of multi-class classification. Nevertheless, their approach had promising results. Their proposed model had an detection accuracy for median filtering (5 × 5 kernel), with input image size 64 × 64 followed by JPEG compression quality factor (QF) 70 and 90 of 94,12% and 96,84% respectively, compared to JPEG compression only. The detection accuracy for median filtering (5 × 5 kernel) with input image size 32 × 32 was 88,65% and 93,21% for JPEG compression quality factor 70 and factor 90, compared to JPEG compression only.

The approach of adding an additional filter to the CNN to suppress the image content was also recognised by Kim and Lee [15]. They proposed a model composed of 1 high pass filter, 2 convolutional layers, 2 max pooling layers and 2 fully connected layers. The output layer used a softmax function to score each class. The high pass filter passes on signals with a frequency higher than a certain cutoff value and attenuates signals with a frequency lower than the cutoff value [16]. The purpose of this High Class Filter (HPF) in the convolutional network is to extract hidden features within the image [15]. The full architecture of the CNN model can be seen in Fig. 3.

The proposed model was trained and tested to identify four different manipulations operations: median filtering (5 × 5), additive white Gaussian noise (AWGN; $\sigma = 2$), Gaussian blurring (5 × 5, $\sigma = 1.1$), and re-sampling (scaling factor 1.5). Their results showed that the initial accuracy of detecting the original image was low but increased as the learning progressed. Their proposed model was able to reach an overall accuracy of 96,67%. The accuracy of the different manipulation operations can be seen in Fig. 7. The results of different numbers of training epochs showed that accuracy does not always increase as learning progresses. For some manipulation operations detection became more accurate, but others decreased in accuracy [15].

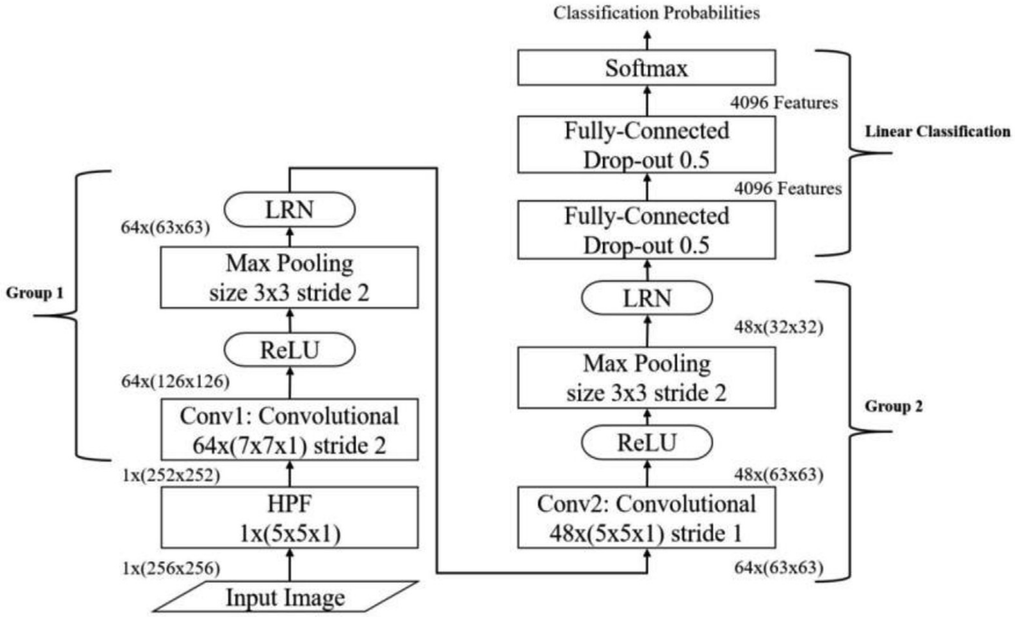

Classification Probabilities
Softmax
4096 Features
Fully-Connected Drop-out 0.5
4096 Features
Fully-Connected Drop-out 0.5
Linear Classification
LRN
48x(32x32)
Max Pooling size 3x3 stride 2
ReLU
48x(63x63)
Conv2: Convolutional 48x(5x5x1) stride 1
64x(63x63)
Group 2
Group 1
LRN
64x(63x63)
Max Pooling size 3x3 stride 2
ReLU
64x(126x126)
Conv1: Convolutional 64x(7x7x1) stride 2
1x(252x252)
HPF 1x(5x5x1)
1x(256x256)
Input Image

Fig. 3. CNN architecture as proposed by Kim and Lee [15] consisting of 1 HPF, 2 convolutional layers, 2 max pooling layers, and 2 fully connected layers with softmax function for classification. The networks input dimension is a 256 × 256 sized grayscale image.

4.2. Constrained convolutional layer

To overcome the need for preliminary feature extraction or preprocessing, Bayar and Stamm [9] proposed a new convolutional layer: the constrained convolutional layer. “The key idea behind developing this layer is that certain local structural relationships exist between pixels independent of an image’s content” [9]. Manipulation of the image will modify these local relationships between pixels in a traceable manner. Consequently, the manipulation detection feature extractors must learn these relationship between a pixel and its neighbouring pixels, while at the same time suppressing the content of the image to prevent the network from learning content dependent features.

To accomplish this Bayar and Stamm [9] developed the constrained convolutional filters that are restrained to learn only a selection of prediction error filters. “Prediction error filters are filters that predict the pixel value at the centre of the filter window, then subtract this central value to produce the prediction error” [9]. The filters in this first constrained convolutional layer is initialised by randomly assigning each a filter weight and subsequently enforce the constraint of the prediction error filters. During training, the constraints are imposed on the filters after each gradient descent update of the filters’ weight [9].

Their full model consisted of 1 constrained convolutional layer, 2 convolutional layers with ReLU activation function and max pooling layer, and 3 fully connected layers with softmax activation function for classification (see Fig. 4). The model was trained and tested to perform universal manipulation detection of four different editing operations: median filtering (5 × 5), Gaussian noise ($\sigma = 2$), Gaussian blurring (5 × 5, $\sigma = 1.1$), and re-sampling (scaling factor 1.5). The input to their model were 227 × 227 sized, grayscale images.

The proposed model was able to achieve an overall accuracy of 99,11% in detecting the four different manipulation operations [9]. The accuracy of the individual manipulation operations using the multi-class classification approach can be seen in Table 5.

Building on their previous research in Ref. [9], Bayar and Stamm [10] performed a series of experiments to systematically examine the influence of several important CNN design choices to guide the architecture of CNN models for image manipulation detection. They investigated (1) the choice of the initial CNN layer, (2) the effect of different types of nonlinearity following the first layer (e.g. pooling, non-linear activation function, etc.), (3) the performance of different pooling techniques (i.e. max pooling and average pooling) (4) the influence of network depth and the effect of integrating a 1 × 1 layer into the CNN to learn associations across feature maps, (5) the influence of the choice of activation function (e.g. ReLU, PReLU etc.), and (6) the effect of different normalisation layers (e.g. BN, LRN) [10].

Their baseline architecture consisted of 1 constrained convolutional layer, 4 convolutional layers of which the first three were followed by a max pooling layer and the fourth by an average pooling layer and 3 fully connected layers with softmax function as can be seen in Fig. 5. The input to their CNN model is the green layer of an image patch sized 256 × 256. The (baseline) CNN architecture is trained to perform universal manipulation detection using five different editing operations: median filtering (5 × 5), Gaussian noise ($\sigma = 2$), Gaussian blurring (5 × 5, $\sigma = 1.1$), re-sampling (scaling factor 1.5)and JPEG compression (QF = 70).

1) *Choice of initial layer.* They considered two alternatives for the initial convolutional layer with the objective to suppress an image’s content and capture pixel values dependencies, the high-pass filter (HPF) [15,17] and the constrained convolutional layer [9]. The CNN model with the constrained convolutional layer outperformed the HPF model with an overall accuracy of 98,70% compared to 97,99%, as can be seen in Table 1. This suggests that the constrained convolutional layer is capable of extracting image manipulation features that may not be captured using a hand-designed HPF [10].
2) *Introducing non-linearity.* They investigated the performance of the proposed model with the introduction of different non-linear operations (i.e. PReLU + max pooling, max pooling and absolute value) following the “constrained convolutional layer”. The overall accuracy per design option can be seen in Table 1. The baseline model without the introduction of any non-

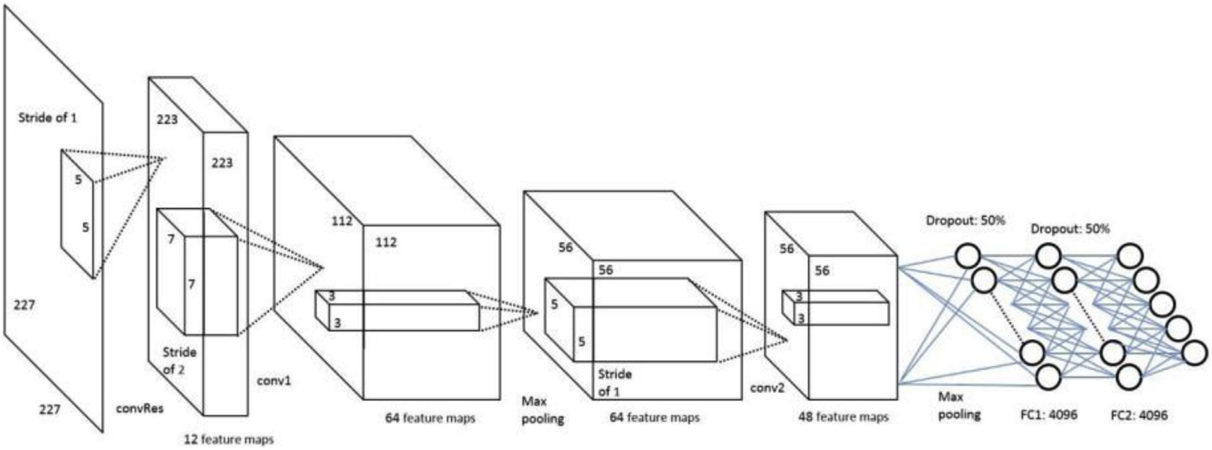


Fig. 4. CNN architecture as proposed by Bayar and Stamm [9] consisting of 1 constrained convolutional layer, 2 convolutional layers, 2 max pooling layers, and 3 fully connected layers with softmax function for classification. The networks input dimension is a 227 × 227 sized grayscale image.

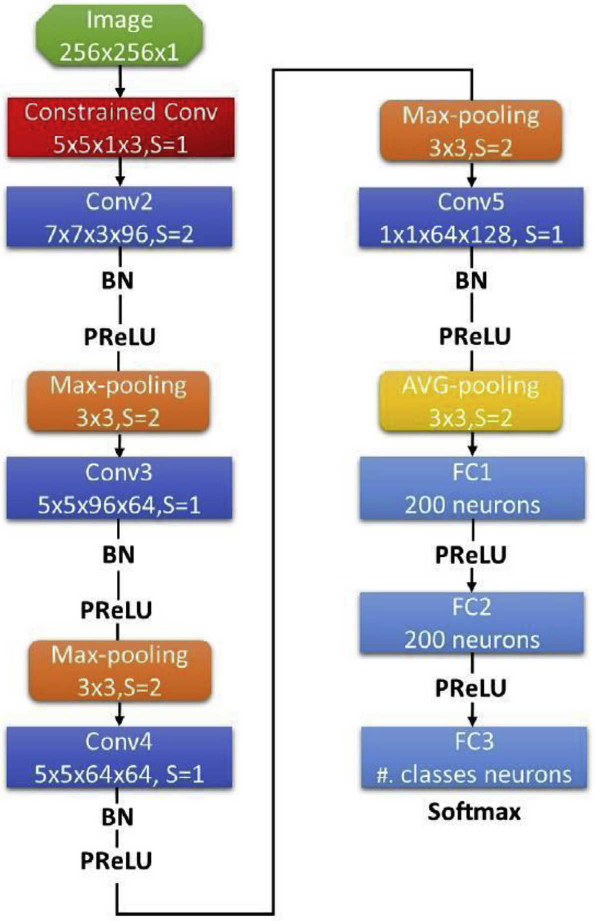


Fig. 5. Baseline CNN architecture as proposed by Bayar and Stamm [10] consisting of 1 constrained convolutional layer, 3 convolutional layers with PReLU activation functions, 2 max pooling layers and 1 average pooling layer, and 3 fully connected layers with softmax function for classification. The networks input dimension is a 256 × 256 green layer image. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

linearity following the constrained convolutional layer performed the best with an accuracy of 98,79%. The results suggest that any type of nonlinearity introduced to the prediction-error features learned by the constrained convolutional layer, inhibits representative features and drops the overall detection rate [10].

3) *Network depth.* Since there is no systematic way to determine the necessary depth in a CNN architecture, Bayar and Stamm [10] assessed the performance of their convolutional neural network with different depths experimentally. They started with one convolutional layer following the constrained convolutional layer and increased the number of convolutional layers while keeping the number of fully connected layers fixed. At every depth the model was trained with and without a 1 × 1 convolutional layer after the last convolutional layer to investigate how important it is to learn association across the feature maps and whether the 1 × 1 convolutional filter could improve final detection rate [10]. The overall detection accuracy for each layer depth with and without 1 × 1 convolutional layer can be seen in Table 1. The results show that with two, three and four convolutional layers, the 1 × 1 convolutional layer improved detection rates [10]. The best performance was achieved when they used three convolutional layers followed by a 1 × 1 convolutional layer (i.e. baseline architecture) with a detection accuracy of 98,70%.
4) *Pooling layer.* According to Bayar and Stamm [10] choosing the correct pooling layer following the 1 × 1 convolutional layer of the baseline architecture is critical for the performance of the CNN. The 1 × 1 filters are capable of learning the association between the highest-level feature maps in the network before they are fed to the fully-connected layers to perform classification. It is important to choose a pooling layer that keeps the most representative features. They compared the performance of an average pooling layer to a max pooling layer following the 1 × 1 convolutional layer. As can be seen in Table 1 using a max pooling layer instead of an average pooling layer decreased detection accuracy from 98,70% to 97,45%. These results suggest that the average pooling layer retains the most representative features from the deepest convolutional feature maps in the network for image manipulation detection.
5) *Activation function.* They compared the performance of baseline architecture with parametric rectified linear unit (PReLU) as activation function to the performance of the baseline architecture with rectified linear unit (ReLU) as activation function and with the exponential linear unit (ELU) as activation function. The PReLU network outperformed the ELU and ReLU networks with a detection accuracy of 98,7% as can be seen in Table 1. The PReLU network performed 0,92% better than the

ReLu network and 0,18% better than the ELU network. Furthermore using PReLU the proposed CNN model reached a higher constant detection rate in fewer number of epochs [10].

6) *Normalisation layer.* Lastly, they trained the baseline architecture with two choices of normalisation layers after each pooling layer, namely batch normalisation (BN) and 5 × 5 local response normalisation (LRN). Again the baseline architecture using batch normalisation outperformed the LRN-based model with a detection accuracy of 98,70% for BN compared to 95,92% for LRN (see Table 1) [10].

In subsequent research Bayar and Stamm [12] further developed their original CNN architecture as proposed in Ref. [9] based on their research in Ref. [10]. Their modified architecture consists of four conceptual blocks: 1) prediction error feature extraction, 2) hierarchical feature extraction, 3) cross feature maps learning and 4) classification (see Fig. 6).

The first block, i.e. *prediction error feature extraction*, consists of a constrained convolutional layer that suppresses the image's content and constraints the CNN to learn the appropriate prediction error features. This layer learns low-level pixel-value dependency traces caused by a specific manipulation operation.

The second block, i.e. *hierarchical feature extraction*, is capable of learning higherlevel prediction error features and consists of 3 consecutive convolutional layers. Each convolutional layer is followed by a batch normalisation (BN) layer, a non-linear activation function (hyperbolic tangent (TanH)) and a max pooling layer.

The *hierarchical feature extraction* block is followed by the *cross feature maps learning* block and consists of one 1 × 1 convolutional layer capable of learning associations across feature maps. Again followed by a BN, activation function (TanH) and average pooling layer.

The final layer, i.e. *classification*, consist of 3 fully-connected layers followed by softmax activation function in the output layer. However, they considered that other options than the softmax function might perform better in the final classification decision. Therefore, they also trained an extremely randomised tree (ET) classifier to calculate the final classification decision. The input to their proposed model is a grayscale image patch, sized 256 × 256.

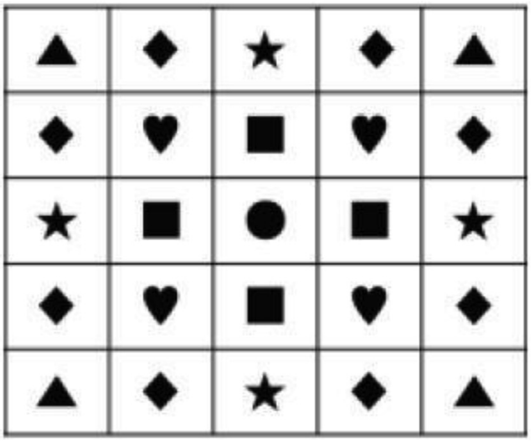


Fig. 7. The constrained weights of a 5 × 5 isotropic filter as proposed by Chen et al. [18].

Compared to the original CNN architecture in Ref. [9], the new CNN architecture accommodates less filters in the constrained convolutional layer, different filters in the third convolutional layer, a different number of filters in the third and fourth convolutional layer, and an additional convolutional layer, and 1 × 1 convolutional layer. Furthermore, the new CNN architecture makes use of average pooling instead of max pooling before the feature output maps are fed to the fully connected layer, it uses different activation functions plus batch normalisation and it consists of a different number of neurons in the fully connected layers.

Their proposed network was trained and tested as universal image manipulation classifier for the detection of five different manipulation operations: median filtering (5 × 5), Gaussian blurring (5 × 5, $\sigma = 1.1$), Gaussian noise ($\sigma = 2$), resampling (scaling factor 1.5) and JPEG compression (QF = 70). The overall manipulation detection accuracy of their proposed model with softmax activation function was 99,26%. The extremely randomised trees

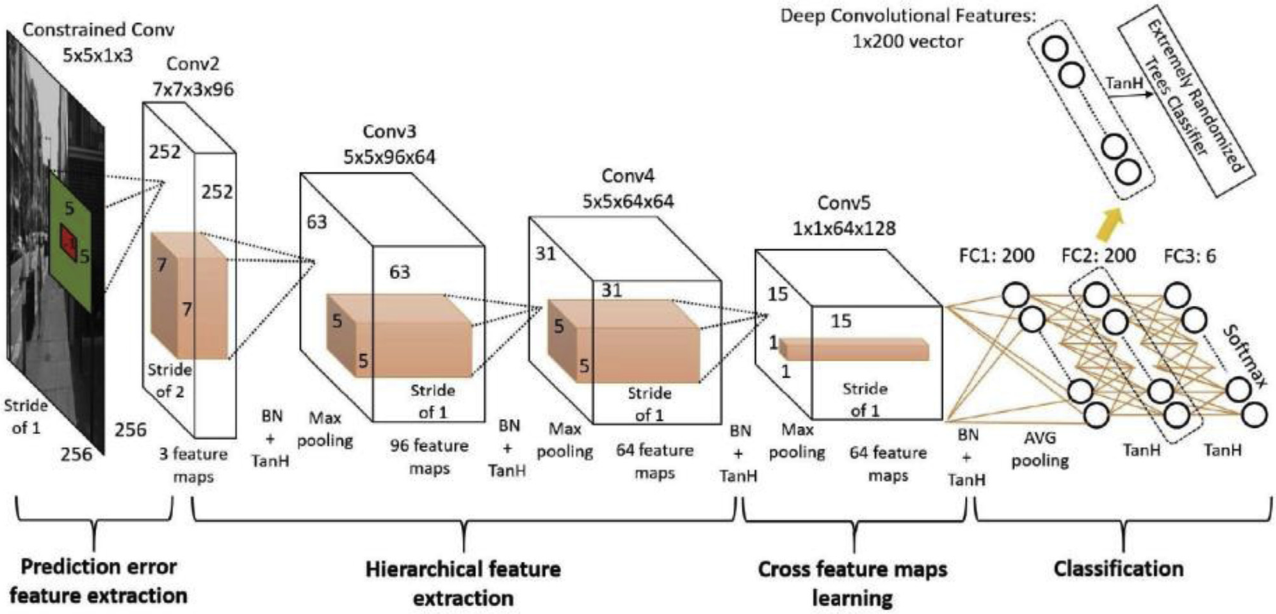


Fig. 6. CNN architecture as proposed by Bayar and Stamm [12] consisting of 1 constrained convolutional layer, 4 convolutional layers, 3 max pooling layers, 1 average pooling layer and 3 fully connected layers with softmax function/extremely randomised tree for classification. The networks input dimension is a 256 × 256 sized grayscale image [12].

classifier increased the overall classification rate to 99,66% [12]. Table 5 shows the detection accuracy of the individual manipulation operations for both the softmax and ET classifier.

The performance of the proposed model was also tested using arbitrary parameters for the five different manipulation operations: median filtering (K_{size}: 3, 5, 7, 9), Gaussian blurring (K_{size}: 3, 5, 7, 9), Gaussian noise ($\sigma = 1.4, 1.6, \ldots, 2$), resampling (scaling factor: 1.2, 1.4, …, 2) and JPEG compression (QF = 60, 61, …, 89, 90). The arbitrary parameter settings are more in line with the realistic scenario where the parameters of the manipulation operations are unknown. The performance of the proposed model for the detection of manipulations operations with fixed parameters compared to arbitrary parameters can be seen in Table 2. Overall the detection accuracy decreased using arbitrary parameters compared to detection accuracy using fixed parameters, but overall detection rates are still high. Manipulation detection using arbitrary parameters did require a larger data set to train the model compared to fixed parameters [12].

As discussed previously different design choices for the convolutional neural network can influence the ultimate image manipulation detection rate. Bayar and Stamm [12] investigated: (1) the choice of the initial CNN layer, (2) the effect of different constrained convolutional layer parameters, (3) the influence of different pooling layers, (4) the performance of different activation function and (5) the effect of the stride size in the second convolutional layer with different input patch sizes on the detection rate of their proposed model.

1) *Choice of initial layer.* They compared the performance of the proposed model using a constrained convolutional layer, no constrained convolutional layer and replacing the constrained convolutional with a generic fixed high-pass filter. The results showed that the overall detection accuracy using the constrained convolutional layer with softmax activation function outperformed the model without constrained convolutional layer and high pass filter with a detection accuracy of 99,26%. The model's performance with no constrained convolutional layer decreased with 0,90% to 98,36% compared to the model with constrained convolutional layer. And the model's performance with the high pass filter decreased with 0,31% to 98,95% compared the constrained convolutional filter.
2) *Different constrained convolutional layer parameters.* Bayar and Stamm [12] varied the number of filters in the constrained convolutional layer from 1 to 6 and subsequently the filter size, using filters of 3 × 3, 5 × 5 and 7 × 7. The CNN's performance maximised when three constrained filters with filter size 5 × 5 were used, with a detection accuracy of 99,26%.
3) *Pooling layer.* They assessed the effect of the pooling layer choice using three different types of pooling layers following the fifth convolutional layer, namely max pooling, average pooling and max-pooling with average pooling. The best identification rates were achieved with max pooling with average pooling after the fifth convolutional layer compared to average pooling only and max pooling only, with an overall accuracy of 99,26%. Moreover, the average pooling layer based CNN converged noticeably slower and to a lower overall accuracy compared to the other two alternatives.
4) *Activation function.* The results of the performance of the proposed model using different activation functions, ELU, ReLU, PReLU and TanH, showed that the TanH activation function had the highest detection accuracy. Furthermore, both TanH and ReLU converged slightly quicker to a higher accuracy compared to ELU, and PReLu [12].
5) *Convolutional stride size.* "The choice of the convolutional stride size is important since it will determine the dimension of features throughout the CNN. The bigger the convolutional stride, the smaller the dimension of the feature maps produced by the CNN" [12]. The detection rate of the proposed CNN using a stride of 1 versus a stride of 2 in the second convolutional layer were compared using different input patch size (64 × 64, 128 × 128 and 256 × 256). For image patches sized 128 × 128 and 64 × 64, a CNN using a stride of 1 outperformed the one using a stride of 2. With patches sized 256 × 256, a CNN with a stride of 2 achieved higher identification rates than the CNN with a stride of 1.

4.3. Isotropic convolutional filter

Regular convolutional neural networks tend to extract features unrelated to the detection of image manipulation [9]. To overcome this problem Chen et al. [18] proposed a convolutional neural network architecture using convolutional layers with an isotropic

Table 1
The overall detection accuracy of the CNN for universal image manipulation detection of median filtering, Gaussian blurring, Gaussian noise, resampling and JPEG compression as proposed by Bayar and Stamm [10] with different design choices.

Design choice	Design choice options	Accuracy
(1)Initial layer	Constrained conv layer (baseline)	**98,70%**
	High pass filter (HPF)	97,99%
(2) Non-linear operation	Without non-linearity (baseline)	**98,70%**
	PReLU + max pooling	95,48%
	Max pooling	93,19%
	Absolute value	94,90%
(3) Network depth	3 conv layers + 1 × 1 conv layer (baseline)	**98,70%**
	1 conv layers + 1 × 1 conv	97,46%
	1 conv layers	98,01%
	2 conv layers + 1 × 1 conv	98,62%
	2 conv layers	98,07%
	3 conv layers	97,50%
	4 conv layers + 1 × 1 conv	98,16%
	4 conv layers	97,52%
(4) Pooling layer	Average pooling (baseline)	**98,70%**
	Max pooling	97,45%
(5) Activation function	PReLU (baseline)	**98,70%**
	ELU	98,52%
	ReLU	97,79%
(6) Normalisation layer	Batch normalisation (BN) (baseline)	**98,70%**
	local response normalisation (LRN)	95,92%

Table 2
Accuracy of identifying the manipulation operations with fixed and arbitrary parameters using the CNN model as proposed by Bayar and Stamm [12] with softmax and extremely randomised tree classification.

	Original	Median filtering	JPEG compression	Gaussian blurring	AWGN	Re-sampling	Average accuracy
Bayar and Stamm (2018) [12] Fixed parameters, Softmax	98,70%	99,08%	99,79%	99,15%	99,96%	98,87%	99,26%
Bayar and Stamm (2018) [12] Arbitrary parameters, Softmax	97,21%	99,01%	99,89%	98,73%	98,93%	99,16%	98,73%
Bayar and Stamm (2018) [12] Fixed parameters, Extremely Randomised Tree	99,49%	99,77%	99,79%	99,46%	99,98%	99,51%	99,66%
Bayar and Stamm (2018) [12] Arbitrary parameters, Extremely Randomised Tree	97,73%	99,59%	99,47%	98,93%	98,61%	99,64%	99,00%

filter [18], called the rotation-invariant CNN. Rotation invariance refers to mapping operations that are identical for the image unimportant of it being in rotation of a multiple of 90° or in mirror symmetry. For most enhancement operations rotation invariance is a general and essential feature. It is therefore an important factor to consider in image manipulation detection [18].

The proposed CNN architecture by Chen et al. [18], is a modification of the model proposed by Bayar and Stamm in Ref. [10]. In the model proposed by Bayar and Stamm [10] a constrained convolutional layer serves as pre-processing layer to adaptively learn pixel value dependencies and to suppress image content. According to Chen et al. [18] the extraction of such dependency features may be effective for the detection of operations that are based on adjacent pixels, such as median filtering, but are not suitable for the detection of histogram alterations related to enhancement operations. Therefore, they propose an isotropic filter layer to suppress image content and to learn useful statistical features to detect enhancement operations.

The isotropic filter is a constrained filter wherein all weights are both centre symmetrical and mirror symmetrical [18]. The weights of a 5 × 5 isotropic filter are illustrated in Fig. 7, the figures with the same shape have similar weight values. So, unimportant of being rotated by a multiple of 90°, the filter will perform the same operation on the image. The isotropic filter thereby serves as an extractor capable of adaptively learning the properties of rotation invariance.

In addition, with the use of isotropic filters the amount of parameters can be significantly reduced. If we take the 5 × 5 filter isotropic filter of Fig. 7, there are only six parameters to be learned which is approximately a quarter of the original filter with ($5 \times 5 =$) 25 learnable parameters.

For their model they replaced all convolutional filters in the model by Bayar and Stamm [2] for isotropic filters. An overview of the full architecture of the rotation-invariant CNN can be seen in Fig. 8. It consists of 6 groups: one preprocessing group with the constrained isotropic convolutional layer, four layer groups (group 2–5) each containing an isotropic convolutional layer, batch normalisation, PReLU activation and pooling, and one classification group (group 6) consisting of three fully-connected layers [18]. The input to their model are grayscale images sized 256 × 256.

The proposed rotation-invariant CNN was trained and tested for the detection of six common enhancement operations: unsharp masking sharpening (UMS) with different settings ($\sigma = 1$, $\lambda = 1.5$; $\sigma = 1.3$, $\lambda = 1$; $\sigma = 0.7$, $\lambda = 1$), Gaussian filtering (5 × 5), median filtering (5 × 5), Gamma correction ($\gamma = 0.5$ and 2), histogram equalisation and S mapping. They compared the detection performance of their proposed model with the model proposed by Bayar and Stamm [10].

The detection rate of the model for the six manipulation operations can be seen in Table 5. The results show that the model by Chen et al. [18] outperformed the model proposed by Bayar and Stamm [10] in detecting the six different manipulation operations. The overall accuracy was 97,77%% and 92,81% for Chen et al. [18] and Bayar and Stamm [10] respectively.

Building on their research in Ref. [18] the model was further developed making use of features from densely connected convolutional neural networks. In densely connected convolutional neural networks each layer is connected to every other layer in a feed-forward fashion [19].

The model proposed by Chen et al. [7] consists of eight layer groups. The first layer group is the isotropic convolutional layer. The second to eight layer groups are traditional convolutional layers. Each convolutional layer is followed by batch normalisation and rectified linear units (ReLU). Furthermore it has 3 transition layers, 3 max pooling layers and 1 fully-connected layer. The general architecture of their proposed CNN model, is illustrated in Fig. 9. The input to their model are 256 × 256 sized, grayscale images.

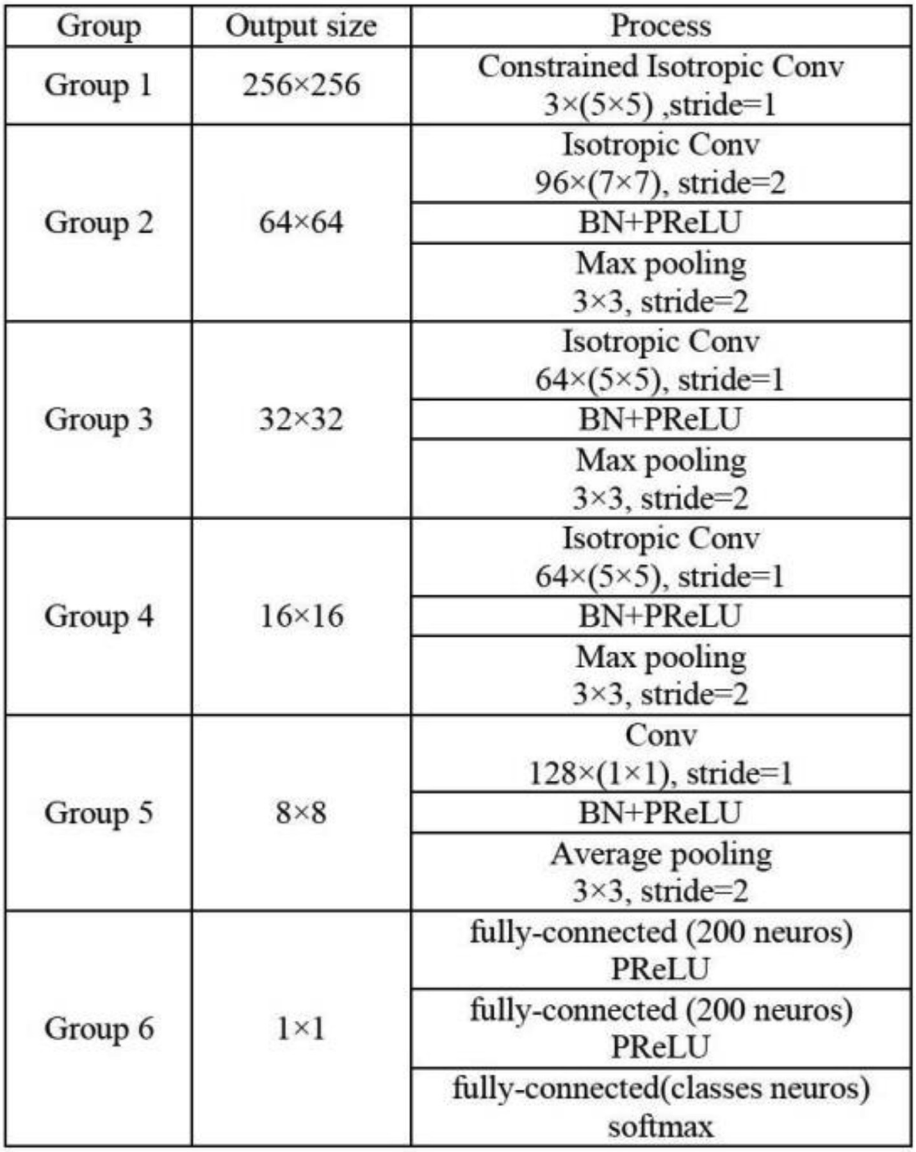
The isotropic filter serves as the extractor that removes anisotropic structures that commonly exist in natural images but are not related to manipulation detection plus it highlights the features that are of interest for forensic analysis. Furthermore, it reduces the number of CNN parameters needed [18]. The transition layers with 1 × 1 convolutions are introduced to lower the number of input feature maps and as a result improve the computational efficiency [7].

With an increase in depth of the CNN, the information extracted in previous layers may have disappeared by the time it reaches the deeper layers. To overcome this problem, Chen et al. [7], make use of the dense connectivity pattern. In dense connectivity two adjacent layers with the same feature map size are connected directly to one another. Compared to the traditional pattern in convolutional neural networks, this dense pattern has better parameter efficiency and it exploits the potential of the network by feature reuse. The proposed model was trained and tested for the detection of five manipulation operations with random parameters and corresponding anti-forensic manipulations shown in Table 3. Anti-forensic operations are techniques used to hide or even remove traces left by image manipulation operations [7].

The detection rate of the individual manipulation operations, including anti-forensic operations can be seen in Table 3. The overall detection rate of their proposed model for classifying multiclass operations was 97,71%.

4.4. No constraints

Experiments with fixed, constrained, and randomly initialised kernels led Boroumand and Fridrich [28] to the notion that no constraints of any kind should be imposed on the filters from the first layer. According to them the fixed or constrained kernels remove information about the image luminance, which can be damaging for example when trying to detect luminance adjustments, such as gamma corrections and brightness and contrast



Group	Output size	Process
Group 1	256×256	Constrained Isotropic Conv 3×(5×5) ,stride=1
Group 2	64×64	Isotropic Conv 96×(7×7), stride=2
		BN+PReLU
		Max pooling 3×3, stride=2
Group 3	32×32	Isotropic Conv 64×(5×5), stride=1
		BN+PReLU
		Max pooling 3×3, stride=2
Group 4	16×16	Isotropic Conv 64×(5×5), stride=1
		BN+PReLU
		Max pooling 3×3, stride=2
Group 5	8×8	Conv 128×(1×1), stride=1
		BN+PReLU
		Average pooling 3×3, stride=2
Group 6	1×1	fully-connected (200 neuros) PReLU
		fully-connected (200 neuros) PReLU
		fully-connected(classes neuros) softmax

Fig. 8. Rotation invariant CNN architecture as proposed by Chen et al. [18] consisting of one constrained isotropic conv layer, 4 isotropic conv layers, 4 pooling layers and three fully connected layers with softmax function.

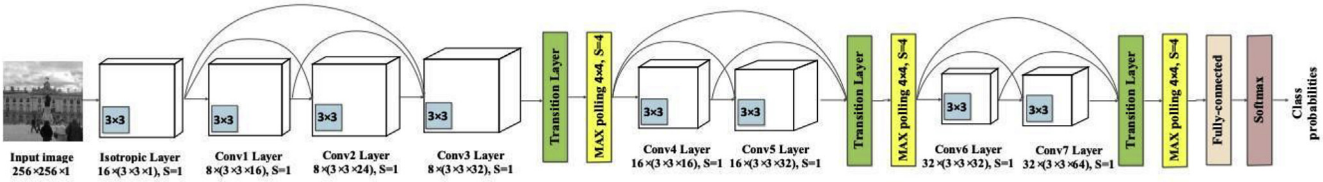


Fig. 9. CNN architecture as proposed by Chen et al. [7] with 8 layer groups, the first being the isotropic convolutional layer and the second to eight the traditional convolutional layer. Additionally, it has 3 transition layers 3 max pooling layers and 1 fully connected layer.

changes. Therefore, their proposed model consists of 8 ‘traditional’ convolutional layers.

Boroumand and Fridrich [28] tested their model with various activation functions and with and without batch normalisation (BN). Their investigation showed the supremacy of the ReLU activation function as well as the benefit of BN that helped speeding up the training performance as well as improving overall performance. Furthermore, they found out that the best performance was obtained by disabling pooling between the first two layers, after which standard 2 × 2 average pooling with stride was applied for each following convolutional layer with the exception of the last where 8 × 8 average pooling layer is applied. The final classification layer consist of one fully connected layer with softmax activation (see Fig. 10). The proposed model is designed for a colour input

Table 3
The Five manipulation classes with random parameter settings and corresponding antiforensic manipulations and the detection accuracy of the multi-class classification model proposed by Chen et al. [7].

Classification class	Parameter	Parameters	Accuracy
Original			93,70%
UMS all	Unsharp masking sharpening (UMS)	σ: 1–1.5, λ: 1–1.5	98.98%
	Anti-UMS [20]	Removing overshoot artefacts in image edges and abrupt change in histogram ends with the same parameter set-up in Ref. [20].	
GC all	Gamma correction (GC)	γ: 0.5, 0.6, 0.7	95.82%
	Anti-GC [21]	Gaussian noise with $\sigma = 1$ is introduced	
	Anti-GC [22]	Adding with random noise of uniform distribution in (−0.5, 0.5)	
MF all	Median filtering (MF)	K_{size}: 3 × 3, 5 × 5, 7 × 7	99,64%
	Anti-MF [23]	Adding with noise disturbance with the same parameter setup in [23]	
	Anti-MF [24]	Adding with random noises with the same parameter setup in [24]	
RES all	resampling (RES)	Random scaling factors: 0.6–2	98.98%
	Anti-RES [25]	Setting the strength of distortion $\sigma = 0.4$	
JPEG all	JPEG compression (JPEG)	Quality factor: 55-95	99.52%
	Anti-JPEG [26]	The original images are JPEG compressed as above, then dither is added in the DCT coefficients	
	Anti-JPEG [27]	The original images are JPEG compressed as above, then modified with [27] corresponding anti-forensic method	

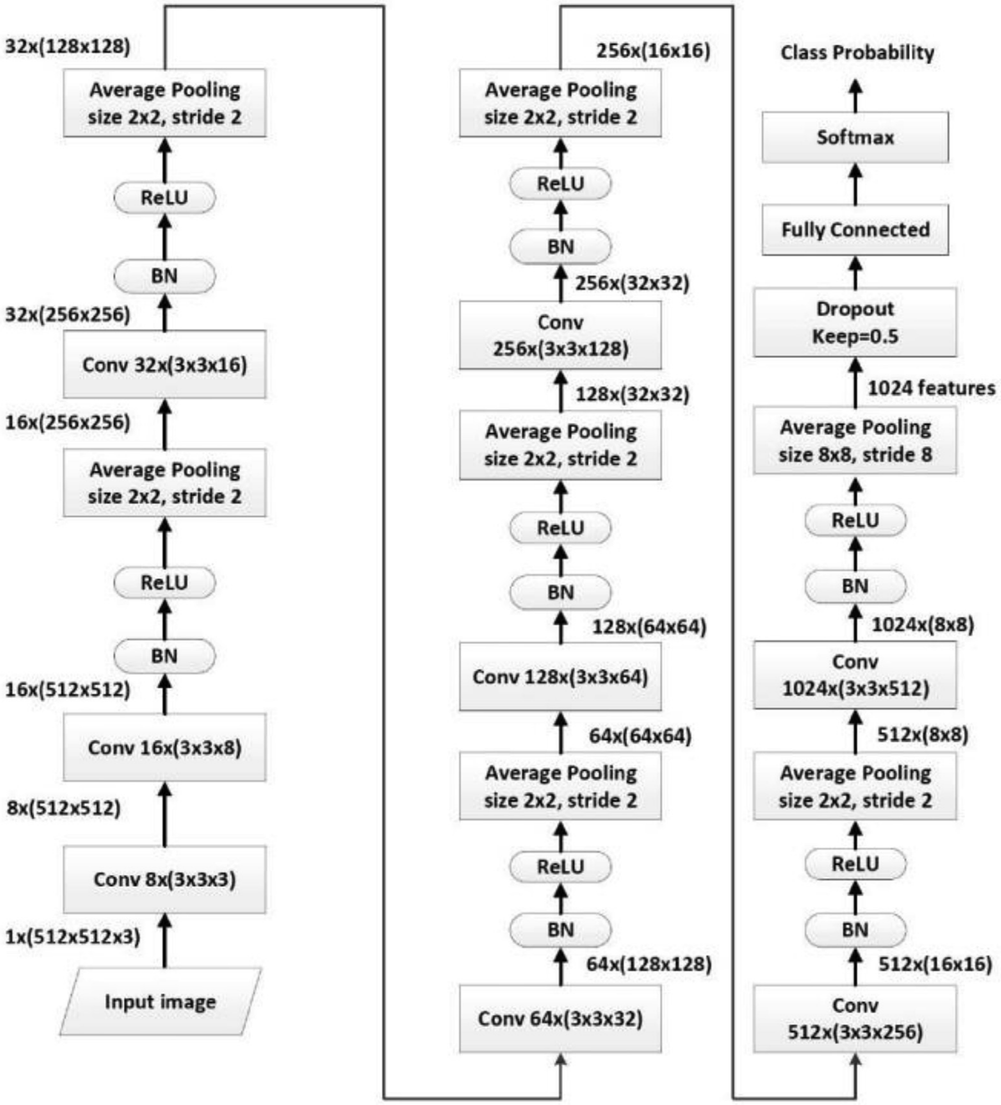


Fig. 10. CNN architecture as proposed by Boroumand and Fridrich [28] consisting of 8 convolutional layers with ReLU activation and batch normalisation, 7 average pooling layers and one fully connected layers with softmax function.

image sized 512 × 512.

The CNN architecture as proposed by Boroumand and Fridrich was trained and tested for the detection of four manipulation classes: low-pass filtering (blurring), high-pass filtering (sharpening), denoising (content adaptive low-pass filtering) and tonal adjustments (histogram equalisation, gamma correction, contrast enhancement etc.). Every manipulation class covered 8 manipulation operations. After applying one of four manipulation class operations, the images were subsequently JPEG compressed with quality factor 85. The performance results per manipulation class can be seen in Table 4. With their proposed architecture they were able to achieve an overall accuracy of 95,22%.

Boroumand and Fridrich [28] also wanted to build a model that is suited for the more realistic scenario wherein images are most likely already JPEG compressed before applying any manipulation, and are saved as JPEG again, after manipulation. While building the CNN-model suited for manipulation detection in the aforementioned scenario Boroumand and Fridrich were faced with two problems. First, the need to consider a range of final JPEG quality factors, rather than a fixed quality factor. And second, the diversification over the downscaling factor that will lead to images with a wide range of sizes, resulting in problems to train the model.

They approached the first problem by training three separate detectors for three different final JPEG quality factors, namely 75, 85 and 95. To solve the second problem, they made two small modification to the CNN architecture described above and trained the network in three separate phases. In phase one the CNN is trained on small images with a fixed size (512 × 512). However instead of computing only the average of each 8 × 8 feature map before they are fed to the fully connected inner-product (IP) layer, they added the minimum, maximum and the variance. Hence, the dimensionality of the input to the fully connected layer becomes 4 × 1024 instead of 1024 [28].

In the second phase the front layer that outputs the 4 × 1024 feature map moments to the fully connected layers is used as a "universal feature extractor" to extract the four statistical moments (i.e. average, minimum, maximum and variance) from all training images. During this phase the model is not trained. The front layer trained in phase 1 is merely used to convert each arbitrarily sized image in the training set to 4 × 1024 moments.

In the third phase two fully connected layers are trained to classify the 4096 (= 4 × 1024) dimensional vectors of moments extracted from all training images. Followed by a softmax function. They believe that the four statistical moments provide the fullyconnected layer with sufficient information to allow the CNN to adjust itself to accurately classify manipulations applied to images of arbitrary size and resolution [28]. For final JPEG re-compression the overall accuracy of the multi-class manipulation detection was 95,84% for QF 75, 97,10% for QF 85 and 97,91% for QF 95.

4.4.1. CNN training

Convolutional neural networks require large amounts of data to train due to their big learning capacity. Zhan et al. [29] present a new approach to train CNN models for multiple-class image manipulations detection using transfer learning.

Table 4
Detection accuracy of the model proposed by Boroumand and Fridrich [28] for original and four different manipulations classes.

	Accuracy
Original	92,3%
Low-pass	96,6%
High-pass	92,6%
Denoising	98,6%
Tonal	95.9%

Traditionally machine learning algorithms use statistical models that are trained on previously collected labelled and unlabelled data to make predictions on future data. Most of these algorithms assume that the distribution of labelled and unlabelled data is the same. However, transfer learning permits for the domains, tasks and distributions of the labelled and unlabelled data for training and testing to differ. Research in transfer learning was inspired by the human ability to apply previously learned knowledge to solve new problems or to come up with even better solutions for existing problems. Similarly, learning the convolutional neural network how to classify median filtered images to help the convolutional neural network classifying average filtered images [29].

They use the CNN architecture proposed by Xu et al. in Ref. [30]. This customised deep CNN model can successfully acquire useful statistical information for steganalysis. Steganalysis is the detection of messages, files, images or video's hidden within another file, image, message or video [31]. The overall architecture consists of one preprocessing layer, six convolutional layers, each followed by an activation function, pooling layer and batch normalisation, and one fully-connected layer with softmax activation function. As input they used 512 × 512 sized, grayscale images.

Zhan et al. [29] applied transfer learning in two application settings, namely transfer between tasks and transfer between databases. For transfer learning between tasks they used the standard transfer learning approach, which is to train the base network (i.e. the steganalysis model) and then to copy the first n layers to the first n layers of the target network (i.e. the convolutional neural network). The remaining layers are then initialised randomly and trained towards the target task. For transfer learning between databases they transferred the parameters of the first and the last $6 - n$ parameters and randomly initialised the first 2 to n layers [29].

They trained and tested their multi-class CNN model for the detection of five different manipulation techniques, i.e. JPEG Compression (QF 70), median filtering (5 × 5), contrast enhancement ($\gamma = 0{,}4$), resampling (scaling factor 1.1) and Guassian noise ($\sigma = 2$). The performance of their best model for multiple classification was able to achieve an overall accuracy of 97,25%. The detection accuracy of the multi-class model for the individual manipulation techniques can be seen in Table 5. They observed from the test results that, with the same learning rate, the accuracy declined on both sides of the vertex. They concluded that, with more transferred layer the specificity of the transferred knowledge constraints the capacity of the convolutional neural network to learn new tasks. And, with fewer transferred layers, the performance will decline due to deficiency of the transferred knowledge. Furthermore, the proposed model with transfer learning converged faster compared to traditional CNN model and had more stable test accuracy during training [29].

With regard to parameter transfer between two databases, the results showed that with fixed parameters (i.e. learning rate = 0) the accuracy reached the peak when the first 2–4 layers were randomly initialised. The accuracy dropped on both sides of the vertex. Again, suggesting that more transferred layers decreases the learning capacity while insufficient transferred layers does not provide enough prior knowledge [29]. Their proposed method is capable of training a convolutional neural network with a just a small amount of data in much less time [29].

Mayar et al. [32] investigated if a convolutional neural network trained for one specific multimedia forensic tasks could be used to extract deep features that are applicable for learning a different task. "Deep features are the neuron responses at a particular layer of the CNN, induced by the feeding forward an image through the

Table 5
Overview of the performance of the proposed CNN architectures for the detection of JPEG compression, resampling and image processing operations. AWGN = Gaussian noise; UMS = unsharp masking sharpening.

	Image input	Original	Median filtering (5 × 5 kernel)	Median filtering (3 × 3 kernel)	JPEG compression	Gaussian blurring (5 × 5 kernel; σ = 1.1)	AWGN (σ = 2)	Re-sampling (scaling factor: 1.5)	Gaussian filtering (5 × 5 kernel)	S Mapping	Histogram equalisation	Gamma correction	UMS	Overall accuracy
Kim and Lee (2017) [15]	256 × 256 Gy-scale image	90.92%	99,45%			97,50%	99,48%	95,98%						96,67%
Bayar and Stamm (2016) [9]	227 × 227 grayscale image	98,40%	98,27%			99,75%	99,77%	99,35%						99,11%
Bayar and Stamm (2017) [10]	256 × 256 green layer image		X		X	X	X	X						98,70%
Bayar and Stamm (2018) [12]														
Soft mm:	256 × 256 grayscale image	98,70%	99,08%		99,79%	99,15%	99,96%	98,87%						99,26%
Bayar and Stamm (2018) [12]														
Extremely Randomised Tree	256 × 256 grayscale image	99,49%	99,77%		99,79%	99,46%	99,98%	99,51%						99,66%
Bayar and Stamm (2017) [10] as cited in [18]	256 × 256 grayscale image		98,91%						98.93	94.25%	97,83%	83,72%	92.7%	92,81%
Chen et al. (2018) [18]	256 × 256 grayscale image		99,98%						99,95%	96,21%	99.22%	95,47%	97,88%	97,77%
Zhan et al. (2017) [29]	512 × 512 grayscale image	98,90%		99,90%	99,90%		99,90%					83,20%		96,36%

network" [32]. Research beyond multimedia forensics has shown that deep features generalise to seemingly unrelated tasks. For example, deep features extracted from a CNN pre-trained for object detection could be used for the training of a scene detection classifier and vice verse [33].

They applied two different approaches for learning deep feature extractors: a transfer learning approach and a multitask learning approach. In the transfer learning approach a convolutional neural network is initially learned for the detection of one specific manipulation operation. The lower layers of the CNN are then frozen. These frozen lower layers are now performing as fixed feature extractors and the upper layers are learned to target a different task. So, the knowledge learned for one task is thus "transferred" to another task. The depth at which the CNN layers were frozen varied during retraining to evaluate the hierarchical nature of feature transference. The layers above the shared depth of the CNN's act as task specific classifiers [32]. Fig. 11a illustrates the transfer learning process with sharing depth up to the first fully connected layer.

To learn a single feature extractor whose output consist of deep features that are highly discriminating for multi-class manipulation detection, Mayar et al. [32] proposed the multitask learning approach. In this approach two (or more) CNNs are trained simultaneously on two (or more) different tasks, at the same time the lower layers of both networks are constrained to learn the same parameter settings (i.e. weights and biases). The layers shared by both networks, form a single, unified, feature extractor for deep features capable of discriminating between two (or more) manipulation detection tasks [32]. A graphical representation of the multitask learning approach is shown in Fig. 11b with sharing depth through to the first fully connected layer. Again the layers aloft the shared lower layers perform as the task specific classifier.

For their experiments Mayar et al. [32] used a model architecture as proposed in Ref. [12] that has proven to be effective at manipulation detection and source camera model identification. The convolutional neural network consists of 1 constrained convolutional layer, 4 convolutional layers and 3 fully connected layers with input image patches sized 256 × 256, green colour channel. Mayar et al. [32] distinguish two tasks: image manipulation detection consisting of 5 different manipulation operations (i.e. median filtering (5 × 5), Gaussian blurring ($\sigma = 1.1$), Gaussian noise ($\sigma = 2$), resampling (SF 1.5) and JPEG compression (QF 70) and source camera model identification of 20 different camera models. A baseline network was trained for both tasks individually to provide for comparison measures [32]. The results for single task accuracy was 99,6% for manipulation detection and 97,5% for camera model identification.

When deep features trained for manipulation detection were transferred to the camera model identification task, the network was able to reach an accuracy of 97,5% if the shallowest share depth was used, consisting of the constrained convolutional layer alone. Accuracy gradually decreased, with increased share depth to 57,8% at the deepest share depth up to the second fully connected layer. When deep features trained for camera model identification were transferred to the manipulation detection task, the network was able to achieve an accuracy of 99,8% at the shallowest shared depth. As the shared depth increased the accuracy of the network decreased, with an accuracy of 97,6% at the deepest shared depth up through fully connected layer 2.

The difference in accuracy drop when we use camera model identification learned features for manipulation detection task compared to when we use manipulation detection learned features

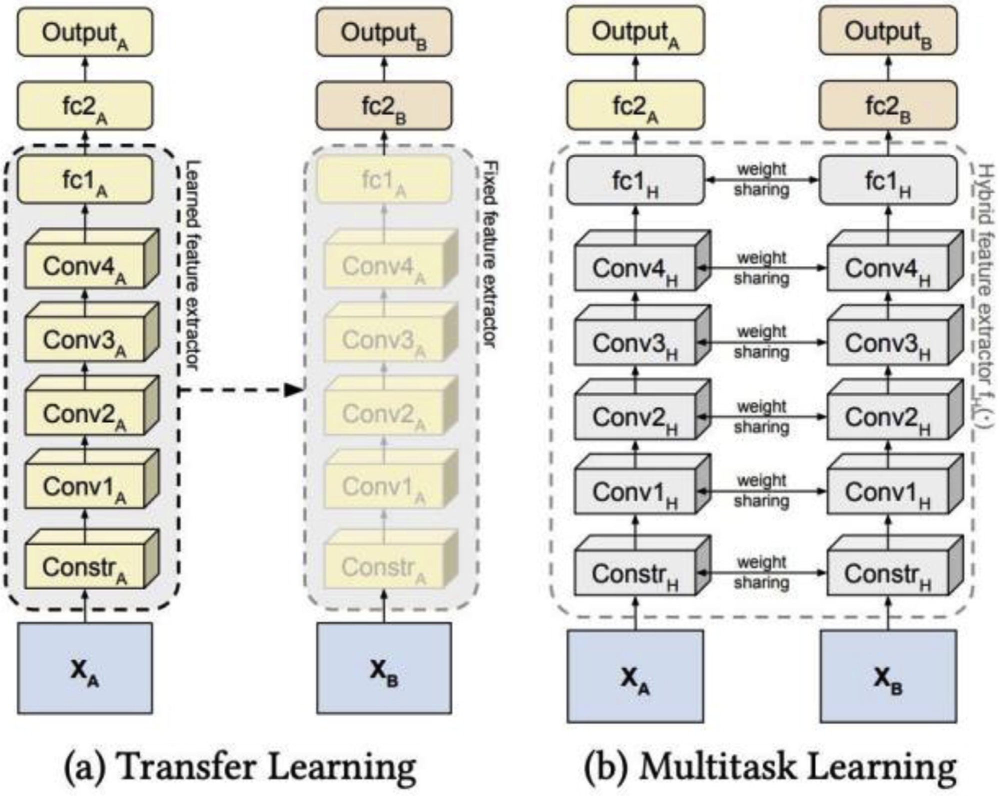


Fig. 11. Graphical representation of proposed approach by Meyer et al. [32] (a) transfer learning and (b) multitask learning, both using an example share depth up to first fully connected layer (fc1).

for camera identification task suggests that there exists a task asymmetry in the generality of forensic deep features. In other words, the transfer of features extracted from the camera models to the manipulation detection task is much better than the transfer of manipulation features to the camera model identification task. A possible explanation could be that camera model features are much more complex than the manipulation features [32]. Furthermore, the decrease in detection accuracy with increasing sharing depth suggests there exists a feature hierarchy. Lower level features learned by the shallower layers are general across tasks, meaning that higher level features can be successfully learned from the low-level representation. The high-level features learned in the deeper layers of the network tend to be more task specific.

When they applied the multitask learning approach the detection accuracy improved at all share depths for the camera model identification task compared to the transfer learning method. At the highest sharing depth up through convolutional layer 2, the multitask learning approach was able to achieve an accuracy of 96,8% for the camera model identification task. That is an improvement of 39.0% over the transfer learning approach. For the manipulation detection task accuracy improved for sharing depths up through convolutional layer 1 and 2 compared to the transfer learning approach, with an accuracy of 99.4% for sharing depth up to convolutional layer 2. Which is an improvement of 1.8%. This shows that the unified features of the multitask learning approach are more effective for discerning multiple forensic tasks than the transfer learning approach, but did not improve over the single-task baseline accuracy. The use of the extremely randomised tree (ERT) classifier instead of the softmax function improved the results in each case slightly [32].

4.4.2. Parameter estimation

An important piece in characterising an image's processing history is to be able to determine specifically how each editing operation was applied. Some of the editing operations that are applied to manipulate an image are parameterised. For instance, a user needs to choose a quality factor when compressing an image or a scaling factor when resizing an image. In some cases estimating the manipulation parameter settings could be useful or even necessary to trace back the processing chain or for the detection of multiple editing operations. Manipulation parameter estimates can also be used to reverse the effects of manipulation or provide an investigator with important information on the original image [2].

The development of parameter estimation algorithms to detect new manipulations or improving upon existing algorithms is both difficult and time consuming. To develop a more generic approach that could be easily adapted to perform parameter estimation of different manipulation operations, Bayar and Stamm [2] proposed a data driven approach capable of directly learning estimators from a labelled data set. Therefore Bayar and Stamm [2] approximately reformulated the manipulation parameter estimation as a classification problem. They divided the manipulation parameter set into different subsets and assigned a classification to each subset. They assume that the investigator has knowledge on the kind of manipulation operation that is applied to the image.

They use a CNN architecture that consist of one constrained convolutional layer, three convolutional layers, each with batch normalisation (BN), TanH (hyperbolic tangent) as non linear activation function and max pooling, one 1 × 1 convolutional layer with average pooling and three fully-connected layers with softmax activation function. Fig. 12 depicts the overall architecture of the CNN. The input to their proposed CNN is a grayscale (or green colour layer) 256 × 256 sized image patch [2]. The performance of their proposed generic approach was trained and tested to detect manipulation parameter settings of four different manipulation operations: resampling, JPEG compression, Gaussian blurring and median filtering.

For JPEG compression and resampling they considered two different scenario's. In the first scenario the investigator estimates the parameter settings from a given known parameter set. In the second scenario, which is a more realistic scenario, the parameter settings are arbitrary and the investigator only knows an upper and lower bound.

Resampling: scaling factor estimation. The fixed known set contained the following scaling factor parameters settings: $\Theta = \{50\%,60\%,70\%, \ldots,150\%\}$. The proposed model was able to achieve an 98,40% estimation accuracy for the detection of the different scaling factor parameter settings. In the second scenario, with only an upper and lower bound on the scaling factor, the parameter set was $\Theta = \{[45\%,155\%]\}$ with the following parameter setting intervals $\Phi = \{[45\%,55\%], \ldots,[145\%,155\%]\}$. On average their approach achieved an 95,45% estimation accuracy, with a higher than 93% estimation accuracy on most scaling intervals. The performance of the CNN decreased with down-scaled images.

JPEG compression: quality factor estimation. The fixed known set contained the following quality factor parameter settings $\Theta = \{50,60,70,80,90\}$. The overall estimation accuracy of their proposed model for the fixed parameter setting was 98,90%. The estimation accuracy decreased when the quality factor was high. In the second scenario, where only the upper and lower bound on the quality factor was known, the parameter set was $\Theta = \{[45\%,100\%]\}$ with $\Phi = \{[45,55), \ldots,[85,95),[95,100]\}$ as the quality factor setting intervals. The overall estimation accuracy was 95,92%. With typically a higher than 94% accuracy for estimating the quality factor interval for most JPEG compressed images.

Median filtering: kernel size estimation. According to Bayar and Stamm [2] forgers typically choose an odd kernel size when applying a median filtering operation to the image. Therefore, they assume that the investigator is aware that the forger used a kernel size value from the fixed set $\Theta = \{3 \times 3,5 \times 5, \ldots,15 \times 15\}$. Their proposed model was capable to achieve an overall accuracy of 99,50% on estimating kernel size.

Gaussian blurring: For Gaussian blurring they also investigated two different scenarios. In the first they used CNN to estimate the Gaussian blurring kernel size with size dependant blur variance. In the second they fixed the kernel size and used the network to identify the blur variance. In both scenario's they used fixed sets. In the first scenario the parameter set consisted of the following kernel sizes $\Theta = \{3 \times 3,7 \times 7,11 \times 11,15 \times 15\}$. The overall detection accuracy for Gaussian blurring kernel size was 99,38%. The detection rate decreased when the standard deviation blur variance was > 2, which is equivalent of choosing a kernel size bigger than 7 × 7. In the second scenario, the parameter set consisted of the blur variance settings: $\Theta = \{1,2,3,4,5\}$. Their proposed model could identify the blur variance with 96,94% accuracy. Similarly, when the standard deviation blur variance was > 2 the estimation accuracy decreased.

4.4.3. Multiple manipulations

In many cases of image manipulation the forger applied more than one manipulation operation to create the forged image, frequently followed by JPEG re-compression. An image that holds numerous manipulations will most likely have different statistical properties for every type of manipulation. Choi et al. [8] were one of the first to test a convolutional neural network for the detection of image manipulation with more than one manipulation operation applied to it.

Their proposed CNN architecture consists of three repeating blocks of two convolutional layers with ReLU activation function followed by one max-pooling layer and three fully-connected

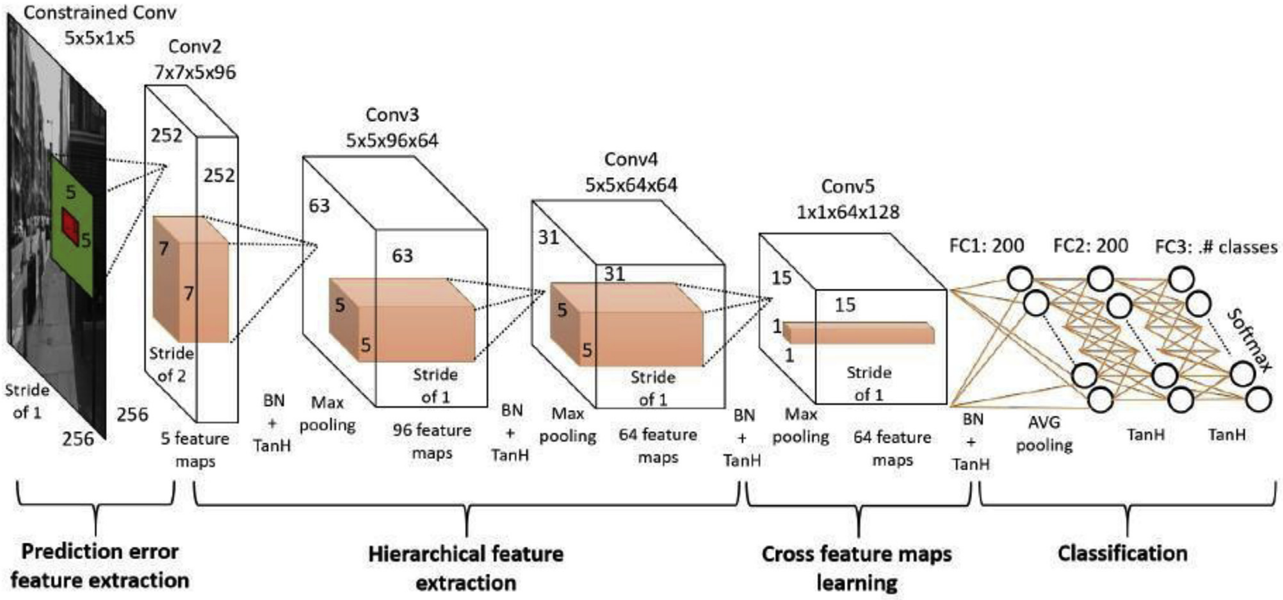


Fig. 12. [2].

layers with softmax activation function. The output layer is a binary classification: manipulated or original image. The input to their model is an RGB3 64 × 64 sized sub-image block. Their proposed method is designed to detect three manipulation operations: Gaussian blurring (3 × 3, $\sigma = 1.1$), median filtering (3 × 3) and Gamma correction ($\gamma = 1/2$) and all its combinations. In addition, the proposed model aimed to detect small sub-image block units, to allow for direct estimation of the operating area in case the image is indeed manipulated.

The performance of the block unit of their proposed architecture can be seen in Table 6. The accuracy for Gamma correction detection was significantly lower compared to the other manipulations, this means that gamma correction detection was not sufficiently trained. Choi et al. [8] also found that the false detection of manipulation operations was predominantly in highly textured regions, defocused regions and very dark regions.

Table 6
Detection accuracy of the model proposed by Choi et al. [8] for original, single and multiple image manipulation operations.

	Accuracy
Original	81,93%
Median filtering (MF)	96,50%
Gaussian blurring (GB)	92,72%
Gamma correction	96,44%
MF-GC	92,85%
GB-GC	92,40%
GB-MF	89,73%
GB-MF-GC	55,91%

4.5. Chain detection

Not only the detection that multiple manipulation techniques were applied is important to determine an image processing history, also the order wherein they were applied can provide the investigator with useful information.

Bayar and Stamm [12] used their CNN model proposed in Ref. [12], to identify an image manipulation history where the image patch was edited by a sequence of up to two different manipulations, and subsequently JPEG compressed (QF 90). The image patches were manipulated using a sequence of the following manipulations: Gaussian blurring ($\sigma = 1.1$, 5 × 5), median filtering (5 × 5) and resizing (scaling factor 1.5). This resulted in the following six combinations of sequences: median filtering-Gaussian blurring, Gaussian blurring-median filtering, median filtering-resizing, resizing-median filtering, resizingGaussian blurring and Gaussian blurring-resizing.

Experiments showed they could reach an overall accuracy of 92,90% with the softmax based CNN and an overall accuracy of 94,19% with the Extremely randomised tree (ET) based CNN. Table 7 shows the performance for the individual manipulations and manipulation chains followed by recompression. Especially the detection rate of the processing operations followed by median filtering were improved by the use of the ERT classifier (see Table 8).

4.5.1. Anti-forensics

In this report we discussed a variety of forensic techniques for the detection of image manipulation [9,10,12,14,15,28]. At the same time farsighted forgers are developing antiforensic techniques [21–27] in an attempt to fool these techniques. Similar to manipulation operation detection techniques most anti-forensic techniques target only one specific type of image anti-forensic. With

Table 7
Detection accuracy of the softmax model compared to the ERT model as proposed by Bayar and Stamm [12] for original, single and multiple image manipulation operations.

	Accuracy Softmax	Accuracy ERT
Original	99,27%	99,33%
Median filtering (MF)	90.54%	91,77%
Gaussian blurring (GB)	93.56%	95,00%
Resampling (RS)	97.15%	98,94%
MF-GB	98,08%	95,87%
GB-MF	80,13%	86,02%
MF-RS	97,69%	99,17%
RS-MF	84,21%	86,00%
GB-RS	93,94%	96,69%
RS-GB	94,50%	93,17%

Table 8
Detection accuracy of the model proposed by Yu et al. [13] for classifying multi-class anti-forensics.

	Accuracy
Original	94,15%
Anti-JPEG [26]	97,1%
Anti-JPEG [27]	99,3%
Anti-median filtering [24]	99,4%
Anti-median filtering [23]	99,5%
Anti-contrast enhancement [22]	91,75%
Anti-resampling [25]	97,35%

convolutional neural networks features for anti-forensic classification can be learned automatically.

Yu et al. [13], developed a 5-layer regular CNN model consisting of four convolutional layers, the second and fourth convolutional layer followed by a max-pooling layers, and one fully connected layer with softmax activation function. There model was trained and tested to detect anti-JPEG compression [26,27], anti-median filtering [23,24], antiresampling [25] and anti-contrast enhancement [22]. According to Yu et al. [13] filter size of the convolutional layer is critical for the network performance. The more suitable the receptive filter the better the resulting extracted features. In their model they use a filter size of 3 × 3 for each receptive field. Furthermore when CNN's are applied for image forensics it is necessary for the pooling units in the pooling layers to overlap to preserve adjacent area's for better performance. An overview of the architecture of the CNN model can be seen in Fig. 13 [13].

To verify if their proposed network was capable of extracting useful and decisive features with increasing depth of convolutional layer, they compared units in the same position of certain feature maps generated by the 4th convolutional layer, between original and anti-forensic images. Yu et al. [13], concluded that it was possible to see perceptible differences between units in the same position, indicating that the network was able to extract useful and decisive features when dealing with the counter ant-forensic task in the training process. In their experiments an architecture with less than four convolutional layers was not capable of extracting a useful feature base [13].

In their experiments the proposed model was able to reach an overall accuracy of 96,96%. The detection accuracy of the individual anti-forensic operations can be seen in Fig. 10.

5. Discussion

The growing interest in the past three years in convolutional neural networks has fuelled research in image manipulation detection and in particular universal image manipulation detection models that are capable of detecting many manipulation operations. The universal manipulation detection approach is less time consuming and does not have the problem of controlling the overall false alarm rate for individual test or handling contradicting outcomes.

Traditional CNN's tend to learn features that capture an image content instead of image manipulation detection features. Researchers have proposed different modifications in the traditional convolutional neural network to suppress the image content and to extract features for manipulation detection, such as a pre-processing layer [14,15], a constrained convolutional layer [9,10,12] and an isotropic convolutional filter [7,18]. All these models were able to achieve an overall accuracy higher than 96% for the detection of median filtering (5 × 5), JPEG compression, Gaussian blurring (5 × 5, $\sigma = 1.1$), Gaussian noise ($\sigma = 2$), Gaussian blurring (5 × 5) and resampling (scaling factor: 1.5) as can be seen in Table 5.

However, according to Boroumand and Fridrich [28] fixed high pas filters or filters constrained to be high pass [9,10], remove important information about the image luminance, which could be harmful for the detection of luminance adjustments, such as gamma correction or brightness and contrast changes. Boroumand and Fridrich compared their model with no additional constraints to the model proposed by Bayar and Stamm [10]. The model by Bayar and Stamm had a lower detection accuracy for all manipulation operation, in particular for gamma correction (see Table 5). Nevertheless, overall accuracy was still above 92%. The high class filter (HPF) as proposed by Kim and Lee [15] has the additional disadvantage that it still requires human intervention to choose a predetermined filter that is not adaptive [10]. This in contrast to the models with a constrained convolution layer, an isotropic filter or with no additional constraints that are capable of extracting all features automatically.

In addition, according to Zhan et al. [29] current preprocessing layers are not able to suppress all the image content. As a consequence the features extracted from images are in general data dependant, which leads to poor generalisation performance when these models are applied to different databases. There are few studies [7,12,18] that tested their model using images of a different database. Their results showed that detection performance slightly decreased. Suggesting that the features learned by the classifier are associated with the training data.

Adding a pre-processing layer can improve the model by actively suppressing an image content, but it might not suppress all

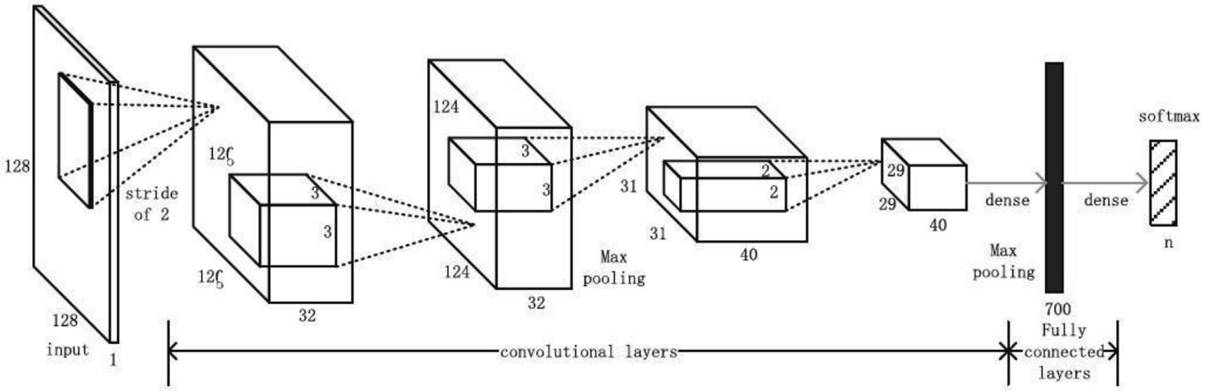


Fig. 13. CNN architecture as proposed by Yu et al. The networks input dimension is a 128 × 128 grayscale image (16.384 neurons). The architecture 5 convolutional layers, two pooling layers and one fully connected layer connected to the output layer through a soft max function [9].

image content. The models proposed by Boroumand and Fridrich [28] and Yu et al. [13] show that it is actually possible to train a traditional convolutional neural network with no constraints to suppress an image content and detect image manipulation by using a small convolutional filter (3 × 3) and disabling pooling between the first and second convolutional layer. Nevertheless, all CNN architectures discussed in this report are aimed at the detection of JPEG compression, resampling and image processing operations. Up to date there is only one paper addressing copy-move and splicing using convolutional neural networks [34]. Because this paper uses a binary classification approach (original/manipulated) instead of a multi-class classification approach this paper is not further discussed in this report.

As mentioned before training a convolutional neural network requires large amounts of data. A CNN's performance depends highly on the size and quality of the training set. For example, using a larger database improves detection accuracy [12]. In the studies discussed in this report the researchers either used self acquired images or an existing database, such as the BOW database, the Boss Base database or the Dresden Image Database to extract authentic images. From these authentic images researchers created their own data set of manipulated images for training and testing. The deep feature approach for transfer and multitask learning as proposed by Mayar et al. [32] could prove to be useful when there is not enough training data to robustly train a full CNN from scratch. It is therefore highly recommended to further investigate the possibilities of transfer learning and multitask learning for convolutional neural network to decrease the need for large amounts of data. Furthermore, it could be useful to develop an open source database containing manipulated image to test and validate CNN models independent of their training database.

Neural networks are developing in a rapid pace. Currently, research in image manipulation detection is already adopting novel deep learning based methods such as deep siamese convolutional neural networks [35], multi-scale convolutional neural networks (MSCNN) [36] and the much faster R–CNN within a two-stream network [37]. Because these are considered extensions of convolutional neural networks they are outside the scope of this report. We see much interest also in deepfake videos that are produced, and the detection of the deepfake videos is getting more difficult since techniques are constantly evolving.

6. Conclusion

In this chapter we discussed the developments in convolutional neural networks for universal manipulation detection, such as the different design and training choices that can be made. The main advantage of convolutional neural networks is that they can automatically learn features for the classification of multiple manipulation operations, without the requirement of human intervention. The results show a high overall detection accuracy (> 92%) for multi-class manipulation detection of JPEG compression, resampling and image processing operations.

The main drawback is the requirement of large amounts of data for training and testing plus the generalisation of the models across databases. Currently there are no publicly available image manipulation databases available for training, testing or validation across models.

Research in image manipulation detection using convolutional neural networks is limited to the detection of manipulation techniques. It is not able to distinguish between "innocent" changing image manipulations, such as red-eye correction, and malicious image manipulations. The proposed models in this report tend to suppress the image content. However understanding the perception of an image content could be very important to distinguish "innocent" from malicious manipulation.

7. Biometric analysis of image material

Biometrics is regularly announced in news items as a panacea against terrorism, security problems, fraud, illegal migration, etcetera. Biometrics, which can be defined as the (automatic) identification or recognition of people based on physiological or behavioral characteristics, is not a single method or technique, but consists of a number of techniques, with each their own advantages and drawbacks. None of the available biometric modalities combines the properties of an ideal biometrics system. We have to acknowledge that biometrics never can be 100% accurate. However, if requirements and applications are carefully considered, biometric systems can provide an important contribution to investigation, authentication and safety.

Within the context of person identification (individualization), different processes can be defined. Within different areas of science, different terminologies are used for the same process, and sometimes the same terminologies are used for different processes. Therefore, a clear definition of the different terms as used in this text is important and made explicit here.

Human Recognition can be defined as the process of identifying or matching a person, his/her photograph or image with a mental image that one has previously stored in long term memory. Recognition requires observation and retention of a person's features and the process of comparison of the retained information with an external image whether it be the life person, a photograph or composite image. The word recognition is important for investigation as well as witness statements. Recognition is within the forensic community also used for the automated searching of a facial image in a biometric database (one-to-many), typically resulting in a group of facial images ranked by computer-evaluated similarity.

Identification is the most contentious term because this most often used term can mean several things in different context, like the automated searching of a facial image in a biometric database (one-to-many) in biometrics, the examination of two facial images or a live subject and a facial image (one-to-one) for the purpose of determining if they represent the same person in forensics, or the assignment of class or family name in biology and chemistry. Therefore, the authors of this paper prefer not to use the term identification unless the meaning is unambiguous within the context.

Recall is here defined as the process of retrieving descriptive information of a person from long term memory in the absence of the person, his/her photograph or other image. Recall requires observation, retention and reproduction of a person's features. Recall is essential for the production of composite images, as produced by a police artist for investigational purposes. However, these images can only be used as investigative tools, and can never be used as proof of identity.

7.1. Pose variation

Pose is the "orientation of the face with respect to the camera, consisting of pitch, roll, and yaw". An optimal frontal pose may be considered as 0° in all directions. Variations to the optimal pose can be due to photographing a physical subject who can move freely during the capture process, or misalignment of the camera. As images are a 2-dimensional representation of the 3-dimensional world, pose of a subject has a major influence on the image captured by a capturing device. As a result of this the appearance and position of facial features can change depending of the pose of the person and the position of the camera at the moment of capture. This is, together with inter and intra observer variability of landmark annotation, one of the main causes of the limited value of landmark measurements on photographs [103]. However, development of pose detection and automatic landmark detection has been reported to result in almost 90% identification accuracy in side view positions [104].

For predicting face recognition performance in a video, it was observed that face detection confidence and face size serve as potentially useful quality measure metrics [105].

7.2. -Dimensional face comparison

The most promising approach to the complicating issues of pose and illumination is the use of 3 dimensional models for pose an illumination correction. Since the previous review, there has been an increase in reports [41–47] on development of methods that are based on the use of 3-dimensional computer models of faces. A number of 3d-acquisition systems are now available for the acquisition of these models. Most 3d-cameras work with a configuration of 1 or more normal digital photo cameras, a flash and the projection of a pattern on the face. These models can be used in two ways. A 3d-facial model of a suspect can be compared to a 3d-model of an unknown person, or the 3d-model of a suspect is used to compute an image that can be compared to an image of an unknown person. Since there are many sources of images and video in practice, a number of studies are focused on the (partial) reconstruction of 3d-models from 1 or more images or video streams. Van Dam et all [106] developed a model 3-D face reconstruction algorithm based on 2D landmarks. he 3D landmark reconstruction algorithm simultaneously estimates the shape, pose and position of the face, based only on the fact that all images in the sequence are recorded using a single calibrated camera.

7.3. Deep learning

With the further development of computer technology, neural network approaches for facial recognition have gained renewed interest. Alignment and the representation of the face by employing explicit 3D face modelling have resulted in improved accuracy of face recognition in unconstrained environments [107–110].

7.4Facial image comparison

The result of facial image recognition is often the selection of 1 or more target facial images that could be matched with the image of the unknown person. In practice, however, this often leads to hit lists with multiple possible matches to the query image, and the correct target not necessarily on top of the hit list. In such cases, the decision has to be made by a forensic anthropologists or forensic image analysts. Since the previous review, more studies and proficiency tests have been reported on the performance of facial image comparison by lay people and experts, showing that there is a reason for concern, and that better methods and technology are needed. A number of institutes have published documents that describe their procedures for performing facial image comparison. These procedures show that measures are being taken to limit the influence of subjective judgments and that there is a need for quantitative statistical data. The FBI has started a working group in 2009 for facial image comparison that is expected to stimulate the development of better methods and technology (FISWG).

Human and computer performance has been systematically compared as part of face recognition competitions, with results being reported for both still and video imagery. Analysis of cross-modal performance shows that for matching frontal faces in still images, algorithms are consistently superior to humans. For video and difficult still face pairs, humans are superior [107].

People doing facial image comparison can be found in four different kinds of professions: forensic photographers, forensic anthropologists, video investigators and imaging scientists. Knowledge of anatomy and physiology of the face is needed to get a good interpretation of differences and similarities in facial features. Similarities or differences in such images can often be explained by differences in the imaging conditions, pointing to the importance of knowledge about optics. Small facial details can be distorted, and artefacts looking like small details introduced due to noise, pixel sampling and compression, requiring knowledge about image processing for the proper interpretation of observations. Changes in image quality, pose and position, lighting and facial expression greatly influence the comparison process. Therefore, it is strongly recommended that one acquire reference images of the suspect and a number of other people with the same video camera in the same situation under similar lighting conditions. While the techniques of facial image comparison are generally accepted within their practitioner communities, they are not tested, and their error rates are unknown. On that basis, the methods of facial image comparison would appear not to meet the anticipated standards [48,109].

It is well-established that matching images of unfamiliar faces is rather error prone. Experimental studies on face matching underestimate its difficulty in real-world situations. Photographs of *unfamiliar* faces seem to be unreliable proofs of identity, especially if the ID documents do not use very recent images of the holders [110].

Existing scientific knowledge of face matching accuracy is based almost exclusively, on people without formal training. Human performance curtails accuracy of face recognition systems, potentially reducing benchmark estimates by 50% in operational settings. Mere practice does not attenuate these limits [111], and some training methods may be inadequate [112]. However, large individual differences have been reported, suggesting that improvements in performance could be made by emphasizing personnel selection [115].

White et al. [114] also have shown that forensic facial examiners outperformed untrained participants and computer algorithms on challenging face matching tests, thereby providing the first evidence that these examiners are experts at this task. Notably, computationally fusing responses of multiple experts produced near perfect performance.

7.5. Eyewitness identification/facial composites

In most of the criminal investigations of a crime, one of the first steps is to interview eyewitnesses. In these interviews the witnesses are asked to provide a description of the perpetrators. For investigational purposes this description may be made into an image by a (police) sketch artist. The sketch artist can also help the witness to recall the face of the perpetrator by showing multiples examples of facial features. Instead of sketches, it is also possible to create photo compositions using examples from databases with facial images.

As not always images of perpetrators are available, matching of composite sketches with facial photographs (e.g. mugshots) is of interest. Matching performance of composite or forensics sketches against photo galleries are promising but still considerably lower than photo matching performance of commercially available systems [117,118].

7.6. Other biometrics

7.6.1. Ear

Even though current ear detection and recognition systems have reached a certain level of maturity, their success is limited to controlled indoor conditions. In addition to variation in illumination, other open research problems include occlusion due to hair, ear symmetry, earprint forensics, ear classification, and ear individuality [119]. The experimental results show that ear recognition may achieve an average rank-one recognition accuracy of more than 95% [120] Current studies are directed towards more robust automated methods for ear detection, landmark localisation and ear recognition using 2D and 3D techniques [121–123].

7.6.2. Body geometry and gait

With the standardisation of photographs, identification primarily occurs from the face. However, results consistently show that less body measurements are needed to find no duplicates when compared to the face. With the combination of eight body measurements, it is possible to achieve results comparable with fingerprint analysis [125]. Thicker garments produce higher inaccuracies in landmark localisation, but errors decrease as placement is repeated. Overall, comparison to truth reveals that on average statures can be predicted with accuracy in excess of 95% [126].

Also lower leg shape, sometimes the only body part consistently depicted in images, has been reported as “an effective biometric trait” [127]. Recent studies have shown that when face identification fails, people rely on the body but are unaware of doing so [128].

Bouchrika et al. [129] reported a method to extract gait features for different camera viewpoints achieving an identity recognition rate of 73.6% processed for 2270 video sequences. Furthermore, experimental results confirmed the potential of the proposed method for identity tracking in real surveillance systems to recognize walking individuals across different views with an average recognition rate of 92.5% for cross-camera matching for two different non-overlapping views.

Yang [130] describes a method for height estimations on eye measurement through a gate cycle.

7.6.3. Soft biometrics

Soft biometric information extracted from a human body (e.g., height, gender, skin colour, hair colour, and so on) is ancillary information easily distinguished at a distance but it is not fully distinctive by itself in recognition tasks. However, this soft information can be explicitly fused with biometric recognition systems to improve the overall recognition when confronting high variability conditions. The use of soft biometric traits is able to improve the performance of face recognition based on sparse representation on real and ideal scenarios by adaptive fusion rules [114]. Depending of the acquisition distance, the discriminative power of the facial regions can change. This results in some cases in better performance than achieved for the full face [131].

Soft biometrics introduce a possibility to automatically search databases based on biometric features obtained from verbal descriptions, resulting in more than 95% identification accuracy [132].

7.6.4. Liveness detection

Spoofing is the act of masquerading as a valid user by falsifying data to gain an illegitimate access. Vulnerability of recognition systems to spoofing attacks (presentation attacks) is still an open security issue in biometrics domain and among all biometric traits. Galbally [133] propose a technique using 25 general image quality features extracted from one image (i.e., the same acquired for authentication purposes) to distinguish between legitimate and impostor samples. The experimental results, obtained on publicly available data sets of fingerprint, iris, and 2D face, show that the proposed method is highly competitive compared with other state-of-the-art approaches and that the analysis of the general image quality of real biometric samples reveals highly valuable information that may be very efficiently used to discriminate them from fake traits. Erdogmus et al. [134] studied detection problem of more complex 3D attack types using various texture based countermeasures.

8. Camera identification of images and video

In criminal investigations of child porn production and distribution, identification of the source of a digital image has become very important, because a specific camera, (or a cell phone camera, a webcam, or a flatbed scanner) could be linked to a suspect using other types of evidence. Identification of images that might have a common source can also be helpful in these investigations. The developments that have been started in the period of the previous review have not been stopped and have lead to a number of new methods and software packages [51–97]. The most used method is based on the estimation of a specific type of fixed pattern noise in an image that is caused by PRNU - *Photo Response Non Uniformity*. The method is also useful in other cases such as murder and fraud to find a links between a camera and images that have been taken.

For identification of a specific camera as the source of a specific image, the PRNU patterns have to be estimated from reference images from the camera and the noise that can be filtered out from this specific image. These patterns have to be compared and a similarity measure is used as a measure for the strength of the evidence that the camera is the source. Common practice is to compare the PRNU pattern of a specific image with the PRNU patterns from a large number of camera’s [51,55,60,61,68–70,75]. The quality of the estimation of the PRNU pattern from an image depends heavily on the image content and this can be taken into account. However, if there are more images available from the same, unknown source, e.g. the frames in a video file [49,50,58,91–95,97,98], much better estimations of the PRNU pattern can be obtained by averaging techniques. In the newer cameras one has to compensate for motion compensation [82,84,88,90]. However several methods are presented to improve the calculation speed as well as clustering images if the camera is not available. Also the use of GPUs is discussed within these methods and optimized with jungle computing [96].

Other sources of fixed pattern noise [52,66,78,85] that have been investigated are based on detection of image artefacts from differences in image processing in the camera chips. Also deep learning is combined with PRNU detection [56,71].

In the forensic practice of a case in which a specific camera has to be identified, a collection of similar cameras from the same brand and type are needed for validation of the results. For using PRNU as evidence, the analyst has to interpret the comparison results. The ENFSI working group [38] for Forensic IT has conducted three proficiency tests to find out what different experts might report to the court about camera identification. In the practice of investigation of large amounts of images, PRNU is also useful to get indications of possibly common sources. A number of studies have been found on the implementation of this application.

The methods are expanded further with the issues of digital zoom as well as with motion compensation algorithms. Furthermore detection of camera model [64,72,80,81] is done, however the forensic usefulness is limited. We also see several papers in the field of manipulation detection [65,72] as well as anti forensics to erase the PRNU pattern and detect this [57,63,73,91].

Disclaimer

This is a republication in journal form of a conference proceeding that was produced for the 19th Interpol Forensic Science Managers Symposium in 2019 and was originally published online at the Interpol website: https://www.interpol.int/content/download/14458/file/Interpol Review Papers 2019.pdf. The publication process was coordinated for the Symposium by the Interpol Organizing Committee and the proceeding was not individually commissioned or externally reviewed by the journal. The article provides a summation of published literature from the previous 3 years (2016-2019) in the field of imaging and video and does not contain any experimental data. Any opinions expressed are solely those of the authors and do not necessarily represent those of their agencies, institutions, governments, Interpol, or the journal.

Declaration of competing interests

The authors have no competing interests to declare.

References

[1] Gajanan K. Birajdar, Vijay H. Mankar, Digital image forgery detection using passive techniques: a survey, Digit. Invest. 10 (3) (2013) 226–245.

[2] Belhassen Bayar, Matthew C. Stamm, A generic approach towards image manipulation parameter estimation using convolutional neural networks, in: Proceedings of 5th ACM Workshop on Information Hiding and Multimedia Security, ACM, Philadelphia, Pennsylvania, USA, 2017, pp. 147–157.

[3] Judith A. Redi, Wiem Taktak, Jean-Luc Dugelay, Digital image forensics: a booklet for beginners, Multimed. Tool. Appl. 51 (1) (2011) 133–162.

[4] Alessandro Piva, An overview on image forensics, ISRN Signal Processing 2013 (2013) 1–22.

[5] Matthew C. Stamm, Min Wu, K.J. Ray Liu, Information forensics: an overview of the first decade, IEEE Access 1 (2013) 167–200.

[6] Mike Nizza, J. Patrick, Lyons, In an Iranian Image, a Missile Too Many, July 2008. (Accessed 16 January 2019).

[7] Yifang Chen, Xiangui Kang, Z. Jane Wang, Qiong Zhan, Densely connected convolutional neural network for multi-purpose image forensics under anti-forensic attacks, in: Proceedings of 6th ACM Workshop on Information Hiding and Multimedia Security, ACM, Innsbruck, Austria, 2018, pp. 91–96.

[8] Hak-Yeol Choi, Han-Ul Jang, Dongkyu Kim, Jeongho Son, Seung-Min Mun, Sunghee Choi, Heung-Kyu Lee, Detecting composite image manipulation based on deep neural networks, in: Proceedings of 2017 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, Poznan, Poland, 2017, pp. 1–5.

[9] Belhassen Bayar, Matthew C. Stamm, A deep learning approach to universal image manipulation detection using a new convolutional layer, in: Proceedings of 4th ACM Workshop on Information Hiding and Multimedia Security, ACM, Vigo, Galicia, Spain, 2016, pp. 5–10.

[10] Belhassen Bayar, Matthew C. Stamm, Design principles of convolutional neural networks for multimedia forensics, Electron. Imag. 2017 (7) (2017) 77–86.

[11] Andrej Karpathy, Lecture Notes Cs231n Convolutional Neural Networks for Visual Recognition, April 2018. (Accessed 15 January 2019).

[12] Belhassen Bayar, Matthew C. Stamm, Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2691–2706.

[13] Jingjing Yu, Yifeng Zhan, Jianhua Yang, Xiangui Kang, A multi-purpose image counter-anti-forensic method using convolutional neural networks, in: Proceedings of 15th International Workshop on Digital Watermarking, 2016, pp. 3–15. Beijing, China.

[14] Jiansheng Chen, Xiangui Kang, Ye Liu, Z. Jane Wang, Median filtering forensics based on convolutional neural networks, IEEE Signal Process. Lett. 22 (11) (2015) 1849–1853.

[15] Dong-Hyun Kim, Hae-Yeoun Lee, Image manipulation detection using convolutional neural network, Int. J. Appl. Eng. Res. 12 (21) (2017) 11640–11646.

[16] Wan Azani Mustafa, Haniza Yazid, Sazali Bin Yaacob, A review: comparison between different type of filtering methods on the contrast variation retinal images, in: Proceedings of 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), IEEE, Batu Ferringhi, Malaysia, 2014, pp. 542–546.

[17] Lionel Pibre, P. Jerome, Dino Ienco, Marc Chaumont, Deep learning for steganalysis is better than a rich model with an ensemble classifier, and is natively robust to the cover source-mismatch, in: Proceedings of Media Watermarking, Security, and Forensics, Part of IS&T International Symposium on Electronic Imaging, EI'2016, IST, San Francisco, California, USA, 2016, pp. 1–10.

[18] Yifang Chen, Zi Xian Lyu, Xiangui Kang, Z. Jane Wang, A rotation-invariant convolutional neural network for image enhancement forensics, in: Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Calgary, AB, Canada, 2018, pp. 2111–2115.

[19] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, Hawai, 2017, pp. 2261–2269.

[20] Laijie Lu, Gaobo Yang, Ming Xia, Anti-forensics for unsharp masking sharpening in digital images, Int. J. Digital Crime Forensics (IJDCF) 5 (3) (2013) 53–65.

[21] Gang Cao, Yao Zhao, Rongrong Ni, Huawei Tian, Anti-forensics of contrast enhancement in digital images, in: Proceedings of 12th ACM Workshop on Multimedia and Security, ACM, Roma, Italy, 2010, pp. 25–34.

[22] Chun-Wing Kwok, Oscar C. Au, Sung-Him Chui, Alternative anti-forensics method for contrast enhancement, in: Proceedings of 10th International Workshop on Digital Watermarking, Springer, Atlantic City, NY, USA, 2011, pp. 398–410.

[23] Zhung-Han Wu, Matthew C. Stamm, K.J. Ray Liu, Anti-forensics of median filtering, in: Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Vancouver, BC, Canada, 2013, pp. 3043–3047.

[24] Duc-Tien Dang-Nguyen, Israel D. Gebru, Valentina Conotter, Giulia Boato, G. Francesco, B. De Natale, Counter-forensics of median filtering, in: Proceedings of 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), IEEE, Pula, Italy, 2013, pp. 260–265.

[25] Matthias Kirchner, Rainer Bohme, Hiding traces of resampling in digital images, IEEE Trans. Inf. Forensics Secur. 3 (4) (2008) 582–592.

[26] Matthew C. Stamm, Steven K. Tjoa, W. Sabrina Lin, K.J. Ray Liu, Antiforensics of jpeg compression, in: Proceedings of 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), IEEE, Dallas, TX, USA, 2010, pp. 1694–1697.

[27] Matthew C. Stamm, K.J. Ray Liu, Anti-forensics of digital image compression, IEEE Trans. Inf. Forensics Secur. 6 (3) (2011) 1050–1065.

[28] Mehdi Boroumand, Jessica Fridrich, Deep learning for detecting processing history of images, Electron. Imag. 2018 (7) (2018) 1–9.

[29] Yifeng Zhan, Yifang Chen, Qiong Zhang, Xiangui Kang, Image forensics based on transfer learning and convolutional neural network, in: Proceedings of 5th ACM Workshop on Information Hiding and Multimedia Security, ACM, Philadelphia, Pennsylvania, USA, 2017, pp. 165–170.

[30] Guanshuo Xu, Han-Zhou Wu, Q. Yun, Shi, Ensemble of cnns for steganalysis: an empirical study, in: Proceedings of 4th ACM Workshop on Information Hiding and Multimedia Security, ACM, Vigo, Galicia, Spain, 2016, pp. 103–107.

[31] Soumyendu Das, Subhendu Das, Bijoy Bandyopadhyay, Sugata Sanyal, Steganography and steganalysis: different approaches, Int. J. Comput. Inf. Technol. Eng. 2 (1) (2008) 1–11.

[32] Mayer Owen, Belhassen Bayar, C. Matthew, Stamm, Learning unified deep-features for multiple forensic tasks, in: Proceedings on 6th ACM Workshop on Information Hiding and Multimedia Security, ACM, Innsbruck, Austria, 2018, pp. 79–84.

[33] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, Aude Oliva, Learning deep features for scene recognition using places database, in: Proceedings of Neural Information Processing Systems 2014, NIPS, Montreal, Canada, 2014, pp. 487–495.

[34] Rao Yuan, Jiangqun Ni, A deep learning approach to detection of splicing and copy-move forgeries in images, in: Proceedings of 2016 IEEE International Workshop on Information Forensics and Security, IEEE, Abu Dhabi, United Arab Emirates, 2016, pp. 1–6.

[35] Aniruddha Mazumdar, Jaya Singh, Yosha Singh Tomar, Prabin Kumar Bora, Universal Image Manipulation Detection Using Deep Siamese Convolutional Neural Network, 2018, pp. 1–6, arXiv preprint arXiv:1808.06323.

[36] Yaqi Liu, Qingxiao Guan, Xianfeng Zhao, Yun Cao, Image forgery localization based on multi-scale convolutional neural networks, in: Proceedings of 6th ACM Workshop on Information Hiding and Multimedia Security, ACM, Innsbruck, Austria, 2018, pp. 85–90.

[37] Peng Zhou, Xintong Han, Vlad I. Morariu, Larry S. Davis, Learning rich features for image manipulation detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, USA, 2018.

[38] Z. Geradts, ENFSI forensic IT working group, Digit. Invest. 8 (Nov. 2011).

[41] Facial comparison overview. https://fiswg.org/documents.html.

[42] E.C. Noyes, Face Recognition in Challenging Situations, PhD thesis, University of York, 2016.

[43] J.P. Davis, K. Lander, R. Evans, A. Jansari, Investigating prdictors of superior face recognition ability in police super-recogniser, Applied Cognitive Psychology, Appl. Cognit. Psychol. 30 (2016) 827–840.

[44] M. Ramon, A.K. Bobak, D. White, Super-recognizers: from the lab to the world and back again, Br. J. Psychol. 110 (3) (2019) 461–479.

[45] P.J. Phillips, A.N. Yates, Y. Hu, C.A. Hahn, et al., Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. www.pnas.org/cgi/doi/10.1073/pnas.1721355115, 2018.

[46] T. Balsdon, S. Summersby, R.I. Kemp, D. White, Improving face identification with specialist teams, Cognit. Res. Princ. Implications 1 (3) (2018), 25-25.

[47] A. Towler, R.I. Kemp, A.M. Burton, J.D. Dunn, T. Wayne, R. Moreton, D. White, Do professional training programs work? PloS One (2019) https://doi.org/10.1371/journal.pone.0211037.

[48] K.L. Ritchie, D. White, R.S.S. Kramer, E. Noyes, R. Jenkins, A.M. Burton, Enhancing CCTV: averages improve face identification from poor-quality images, Appl. Cognit. Psychol. 32 (2018) 671–680, https://doi.org/10.1002/acp.3449.

[49] Aghamaleki Abbasi, Javad, Alireza Behrad, Inter-frame video forgery detection and localization using intrinsic effects of double compression on quantization errors of video coding, Signal Process. Image Commun. 47

(2016) 289−302, https://doi.org/10.1016/j.image.2016.07.001.

[50] Javad Abbasi Aghamaleki, Alireza Behrad, Detecting double compressed MPEG videos with the same quantization matrix and synchronized group of pictures structure, J. Electron. Imag. 27 (1) (2018), https://doi.org/10.1117/1.JEI.27.1.013031, 013031−013031.

[51] K.R. Akshatha, et al., Digital camera identification using PRNU: a feature based approach, Digit. Invest. 19 (2016) 69−77, https://doi.org/10.1016/j.diin.2016.10.002.

[52] Mustafa Al-Ani, Fouad Khelifi, On the SPN estimation in image forensics: a systematic empirical evaluation, IEEE Trans. Inf. Forensics Secur. 12 (5) (2017) 1067−1081, https://doi.org/10.1109/TIFS.2016.2640938.

[53] Irene Amerini, et al., Dealing with video source identification in social networks, Signal Process. Image Commun. 57 (2017) 1−7, https://doi.org/10.1016/j.image.2017.04.009.

[55] Armas Vega, Esteban Alejandro, et al., Digital images authentication technique based on DWT, DCT and local binary patterns, Sensors 18 (10) (2018), https://doi.org/10.3390/s18103372.

[56] Eleni Athanasiadou, et al., Camera recognition with deep learning, Forensic Sci. Res 3 (3) (2018) 210−218, https://doi.org/10.1080/20961790.2018.1485198.

[57] Mauro Barni, et al., An improved statistic for the pooled triangle test against PRNU-copy attack 25 (2018) 10, https://doi.org/10.1109/LSP.2018.2863045.

[58] Blind Detection and Localization of Video Temporal Splicing Exploiting Sensor-Based Footprints, 2018, pp. 1362−1366, 2018-, Eurasip.

[59] Christina Boididou, et al., Verifying information with multimedia content on twitter, Multimed. Tool. Appl. 77 (12) (2018) 15545−15571, https://doi.org/10.1007/s11042-017-5132-9.

[60] Davide Cozzolino, Luisa Verdoliva, Noiseprint: a CNN-based camera model fingerprint, IEEE Trans. Inf. Forensics Secur. (2019), https://doi.org/10.1109/TIFS.2019.2916364, 1−1.

[61] Fixed Pattern Noise Pixel-Wise Linear Correction for Crime Scene Imaging CMOS Sensor vol. 10198 (2017). Spie, 1019802−1019802-14.

[62] Fooling PRNU-Based Detectors through Convolutional Neural Networks, 2018, pp. 957−961, 2018-, Eurasip.

[63] García Villalba, Luis Javier, et al., A PRNU-based counter-forensic method to manipulate smartphone image source identification techniques, Future Generat. Comput. Syst. 76 (2017) 418−427, https://doi.org/10.1016/j.future.2016.11.007.

[64] García Villalba, Luis Javier, et al., Identification of smartphone brand and model via forensic video analysis, Expert Syst. Appl. 55 (2016) 59−69, https://doi.org/10.1016/j.eswa.2016.01.025.

[65] Sonja Georgievska, et al., Clustering image noise patterns by embedding and visualization for common source camera detection, Digit. Invest. 23 (2017) 22−30, https://doi.org/10.1016/j.diin.2017.08.005.

[66] Z.J. Geradts, et al., The use of photo response non-uniformity (PRNU) patterns for the comparison of online videos on social media, Proc. Am. Acad. Forensic Sci. 22 (2016).

[67] Bhupendra Gupta, Mayank Tiwari, An empirical cross-validation of denoising filters for PRNU extraction, Forensic Sci. Int. 292 (2018) 110−114, https://doi.org/10.1016/j.forsciint.2018.09.017.

[68] Bhupendra Gupta, Mayank Tiwari, Improving performance of source-camera identification by suppressing peaks and eliminating low-frequency defects of reference SPN, IEEE Signal Process. Lett. 25 (9) (2018) 1340−1343, https://doi.org/10.1109/LSP.2018.2857223.

[69] Bhupendra Gupta, Mayank Tiwari, Improving source camera identification performance using DCT based image frequency components dependent sensor pattern noise extraction method, Digit. Invest. 24 (2018) 121−127, https://doi.org/10.1016/j.diin.2018.02.003.

[70] Jong-Uk Hou, Heung-Kyu Lee, Detection of hue modification using photo response nonuniformity, IEEE Trans. Circ. Syst. Video Technol. 27 (8) (2017) 1826−1832, https://doi.org/10.1109/TCSVT.2016.2539828.

[71] Na Huang, et al., Identification of the source camera of images based on convolutional neural network, Digit. Invest. 26 (2018) 72−80, https://doi.org/10.1016/j.diin.2018.08.001.

[72] Mehdi Jahanirad, et al., An evolution of image source camera attribution approaches, Forensic Sci. Int. 262 (2016) 242−275, https://doi.org/10.1016/j.forsciint.2016.03.035.

[73] P. Korus, Jw Huang, Multi-scale Analysis strategies in PRNU-based tampering localization, IEEE Trans. Inf. Forensics Secur. 12 (4) (2017) 809−824, https://doi.org/10.1109/TIFS.2016.2636089.

[74] Emmanuel Kiegaing Kouokam, Ahmet Emir Dirik, PRNU-based source device attribution for YouTube videos, Digit. Invest. 29 (2019) 91−100, https://doi.org/10.1016/j.diin.2019.03.005.

[75] A. Lawgaly, F. Khelifi, Sensor pattern noise estimation based on improved locally adaptive DCT filtering and weighted averaging for source camera identification and verification, IEEE Trans. Inf. Forensics Secur. 12 (2) (2017) 392−404, https://doi.org/10.1109/TIFS.2016.2620280.

[76] Jian Li, et al., Extraction of PRNU noise from partly decoded video, J. Vis. Commun. Image Represent. 57 (2018) 183−191, https://doi.org/10.1016/j.jvcir.2018.10.023.

[77] [a] Ruizhe Li, et al., Inference of a compact representation of sensor fingerprint for source camera identification, Pattern Recogn. 74 (2018) 556−567, https://doi.org/10.1016/j.patcog.2017.09.027;
[b] Xiang Lin, et al., Recent advances in passive digital image security forensics: a brief review, Engineering 4 (1) (2018) 29−39, https://doi.org/10.1016/j.eng.2018.02.008.

[78] [a] Xf Lin, Ct Li, Preprocessing reference sensor pattern noise via spectrum equalization, IEEE Trans. Inf. Forensics Secur. 11 (1) (2016) 126−140, https://doi.org/10.1109/TIFS.2015.2478748;
[b] Min Long, et al., Identifying natural images and computer generated graphics based on binary similarity measures of PRNU, Multimed. Tool. Appl. 78 (1) (2019) 489−506, https://doi.org/10.1007/s11042-017-5101-3.

[79] F. Marra, et al., Blind PRNU-based image clustering for source identification, IEEE Trans. Inf. Forensics Secur. 12 (9) (2017) 2197−2211, https://doi.org/10.1109/TIFS.2017.2701335.

[80] F. Marra, et al., On the vulnerability of deep learning to adversarial attacks for camera model identification, Signal Process. Image Commun. 65 (2018) 240−248, https://doi.org/10.1016/j.image.2018.04.007.

[81] Francesco Marra, et al., A deep learning approach for Iris sensor model identification, Pattern Recogn. Lett. 113 (2018) 46−53, https://doi.org/10.1016/j.patrec.2017.04.010.

[82] R. Matthews, et al., An analysis of optical contributions to a photo-sensor's ballistic fingerprints, Digit. Invest. 28 (2019) 139−145, https://doi.org/10.1016/j.diin.2019.02.002.

[83] Richard Matthews, et al., Isolating lens effects from source camera identification using sensor pattern noise, Aust. J. Forensic Sci. (2019) 1−4, https://doi.org/10.1080/00450618.2019.1569133.

[84] Ambuj Mehrish, et al., Robust PRNU estimation from probabilistic raw measurements, Signal Process. Image Commun. 66 (2018) 30−41, https://doi.org/10.1016/j.image.2018.04.013.

[85] Christiaan Meij, Zeno Geradts, Source camera identification using photo response non-uniformity on WhatsApp, Digit. Invest. 24 (2018) 142−154, https://doi.org/10.1016/j.diin.2018.02.005.

[86] Arjan Mieremet, Camera-identification and common-source identification: the correlation values of mismatches, Forensic Sci. Int. 301 (2019) 46−54, https://doi.org/10.1016/j.forsciint.2019.05.008.

[87] Manoranjan Mohanty, et al., E-PRNU: encrypted domain PRNU-based camera attribution for preserving privacy, IEEE Trans. Dependable Secure Comput. (99) (2019), https://doi.org/10.1109/TDSC.2019.2892448, 1−1.

[88] Tong Qiao, et al., Individual camera device identification from JPEG images, Signal Process. Image Commun. 52 (C) (2017) 74−86, https://doi.org/10.1016/j.image.2016.12.011.

[89] S. Saito, et al., A theoretical framework for estimating false acceptance rate of PRNU-based camera identification, IEEE Trans. Inf. Forensics Secur. 12 (9) (2017) 2026−2035, https://doi.org/10.1109/TIFS.2017.2692683.

[90] S. Taspinar, et al., PRNU-based camera attribution from multiple seam-carved images, IEEE Trans. Inf. Forensics Secur. 12 (12) (2017) 3065−3080, https://doi.org/10.1109/TIFS.2017.2737961.

[91] Thanh Hai Thai, et al., Camera model identification based on the generalized noise model in natural images, Digit. Signal Process. 48 (2016) 285−297, https://doi.org/10.1016/j.dsp.2015.10.002.

[92] Mayank Tiwari, Bhupendra Gupta, Image features dependant correlation-weighting function for efficient PRNU based source camera identification, Forensic Sci. Int. 285 (2018) 111, https://doi.org/10.1016/j.forsciint.2018.02.005.

[93] Mayank Tiwari, Bhupendra Gupta, Image features dependant correlation-weighting function for efficient PRNU based source camera identification, Forensic Sci. Int. 285 (2018) 111−120, https://doi.org/10.1016/j.forsciint.2018.02.005.

[94] D. Valsesia, et al., User authentication via PRNU-based physical unclonable functions, IEEE Trans. Inf. Forensics Secur. 12 (8) (2017) 1941−1956, https://doi.org/10.1109/TIFS.2017.2697402.

[95] Diego Valsesia, et al., ToothPic: camera-based image retrieval on large scales, IEEE MultiMedia (99) (2018), https://doi.org/10.1109/MMUL.2018.2873845, 1−1.

[96] B. Van Werkhoven, et al., A jungle computing approach to common image source identification in large collections of images, Digit. Invest. 27 (2018) 3−16, https://doi.org/10.1016/j.diin.2018.09.002.

[97] Venkata, Udaya Sameer, Ruchira Naskar, Eliminating the effects of illumination condition in feature based camera model identification, J. Vis. Commun. Image Represent. 52 (2018) 24−32, https://doi.org/10.1016/j.jvcir.2018.01.015.

[98] Bo Wang, et al., Source camera model identification based on convolutional neural networks with local binary patterns coding, Signal Process. Image Commun. 68 (2018) 162−168, https://doi.org/10.1016/j.image.2018.08.001.

[103] M. Cummaudo, M. Guerzoni, L. Marasciuolo, D. Gibelli, A. Cigada, Z. Obertovà, M. Ratnayake, P. Poppa, P. Gabriel, S. Ritz-Timme, C. Cattaneo, Pitfalls at the root of facial assessment on photographs: a quantitative study of accuracy in positioning facial landmarks, Int. J. Leg. Med. 127 (May 2013) 699−706.

[104] P. Santemiz, O. Aran, M. Saraclar, L. Akarun, Automatic sign segmentation from continuous signing via multiple sequence alignment,, in: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, 2009, pp. 2001−2008.

[105] Y. Lee, P.J. Phillips, J.J. Filliben, J.R. Beveridge, Identifying Face Quality and Factor Measures for Video, 2014. NISTIR 8004.

[106] Landmarkbased modelfree 3D face shape reconstruction from video sequences, in: C. van Dam, R. Veldhuis (Eds.), Proceedings of the 2nd International Business and System Conference BSC 2013, 01-Sep-2013.

[107] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to

human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1701–1708.

[108] Y. Taigman, M.A. Ranzato, T. Aviv, M. Park, DeepFace: closing the gap to human-level performance in face verification, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1–8. Columbus.

[109] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A Unified Embedding for Face Recognition and Clustering, 2015 arXiv preprint arXiv: 1503.03832.

[110] E. Zhou, Z. Cao (Eds.), Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not?, 2015 arXiv: 1501.04690.

[111] P.J. Phillips, A.J. O'Toole, Comparison of human and computer performance across face recognition experiments, Image Vis Comput. 32 (1) (Dec. 2013) 74–85.

[112] X. Mallett, M.P. Evison, Forensic facial comparison: issues of admissibility in the development of novel analytical technique, J. Forensic Sci. 58 (Jul. 2013) 859–865.

[114] D. White, J.D. Dunn, A.C. Schmid, R.I. Kemp, Error rates in users of automatic face recognition software, PloS One 10 (Oct. 2015).

[115] P. Tome, J. Fierrez, R. Vera-Rodriguez, M. Nixon, Soft biometrics and their application in person recognition at a distance, IEEE Trans. Inf. Forensics Secur. 9 (3) (2014) 464–475, 1–1.

[117] D. White, P.J. Phillips, C.A. Hahn, M. Hill, A.J. O'Toole, Perceptual expertise in forensic facial image comparison, Proc. Biol. Sci./R. Soc. 282 (Sep. 2015).

[118] B.F. Klare, K. Bonnen, A.K. Jain, Hu Han, Matching composite sketches to face photos: a component-based approach, IEEE Trans. Inf. Forensics Secur. 8 (Jan. 2013).

[119] S.J. Klum, H. Han, B.F. Klare, A.K. Jain, The FaceSketchID system: matching facial composites to mugshots, IEEE Trans. Inf. Forensics Secur. 9 (Dec. 2014).

[120] A. Abaza, A. Ross, C. Hebert, M.A.F. Harrison, M.S. Nixon, A survey on ear biometrics, ACM Comput. Surv. 45 (Feb. 2013).

[121] A. Kumar, C. Wu, Automated human identification using ear imaging, Pattern Recogn. 45 (Mar. 2012).

[122] A. Pflug, C. Busch, Ear biometrics: a survey of detection, feature extraction and recognition methods, IET Biom. 1 (2012).

[123] A. Pflug, J. Wagner, C. Rathgeb, C. Busch, Impact of severe signal degradation on ear recognition performance, in: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 1342–1347.

[125] T. Lucas, M. Henneberg, Comparing the face to the body, which is better for identification? Int. J. Leg. Med. 130 (Mar. 2016) 533–540.

[126] T. Scoleri, T. Lucas, M. Henneberg, Effects of garments on photo-anthropometry of body parts: application to stature estimation, Forensic Sci. Int. 237 (Apr. 2014) 148.e1–148.e12.

[127] M.R. Islam, F.K.-S. Chan, A.W.-K. Kong, A preliminary study of lower leg geometry as a soft biometric trait for forensic investigation, in: 2014 22nd International Conference on Pattern Recognition, 2014, pp. 427–431.

[128] A. Rice, P.J. Phillips, V. Natu, X. An, A.J. O'Toole, Unaware person recognition from the body when face identification fails, Psychol. Sci. 24 (Nov. 2013) 2235–2243.

[129] I. Bouchrika, J.N. Carter, M.S. Nixon, Towards automated visual surveillance using gait for identity recognition and tracking across multiple non-intersecting cameras, Multimed. Tool. Appl. 75 (2) (Nov. 2014) 1201–1222.

[130] S.X.M. Yang, P.K. Larsen, T. Alkjær, B. Juul-Kristensen, E.B. Simonsen, N. Lynnerup, Height estimations based on eye measurements throughout a gait cycle, Forensic Sci. Int. 236 (Mar. 2014) 170–174.

[131] P. Tome, R. Vera-Rodriguez, J. Fierrez, J. Ortega-Garcia, Facial soft biometric features for forensic face recognition, Forensic Sci. Int. 257 (Dec. 2015) 271–284.

[132] D.A. Reid, M.S. Nixon, Human identification using facial comparative descriptions, in: 2013 International Conference on Biometrics (ICB), 2013, pp. 1–7.

[133] J. Galbally, S. Marcel, J. Fierrez, Image quality assessment for fake biometric detection: application to Iris, fingerprint, and face recognition, IEEE Trans. Image Process.:A Publ. IEEE Signal Process. Soc.y 23 (Feb. 2014) 710–724.

[134] N. Erdogmus, S. Marcel, Spoofing face recognition with 3D masks, IEEE Trans. Inf. Forensics Secur. 9 (Jul. 2014).

Further reading

[39] ISO/IEC 2382-37:2012(E) Information Technology-Vocabulary Part 37, 2012. Biometrics.

[40] P. Grother, M. Ngan, K. Hanaoka, ongoing face recognition vendor test (FRVT) Part 2: identification, Inside NIST 8238 (2018). https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8238.pdf.

[99] A. Dutta, Predicting Performance of a Face Recognition System Based on Image Quality, University of Twente, Enschede, The Netherlands, 24-Apr.

[113] A.M. Megreya, A. Sandford, A.M. Burton, Matching face images taken on the same day or months apart: the limitations of photo ID: matching face images, Appl. Cognit. Psychol. 27 (6) (Oct. 2013) 700–706, n/a–n/a.