UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

VICTOR CHITOLINA SCHETINGER

Beyond Digital, Images, and Forensics: Towards a Regulation of Trust in Multimedia Communication

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Advisor: Prof. Dr. Manuel Menezes de Oliveira Neto

Porto Alegre
August 2018

CIP — CATALOGING-IN-PUBLICATION

Schetinger, Victor Chitolina

Beyond Digital, Images, and Forensics: Towards a Regulation of Trust in Multimedia Communication / Victor Chitolina Schetinger. – Porto Alegre: PPGC da UFRGS, 2018.

164 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2018. Advisor: Manuel Menezes de Oliveira Neto.

1. Digital Image Forensics. 2. Image Composition. 3. Automated Planning. 4. Human-based computation. I. de Oliveira Neto, Manuel Menezes. II. Título.

ABSTRACT

This thesis discusses the role of Digital Image Forensics as a regulator of digital media in society. This includes a perceptual study with over 400 subjects to assess their ability to notice editing in images. The results of such experiment indicate that humans are easily fooled by digital images, not being able to tell apart edited and pristine images. The thesis then analyzes the effectiveness of the available arsenal of digital image forensics technology to detect image editing performed by state-of-the-art image-compositing techniques. By analyzing fundamental image patterns, forensics techniques can effectively detect the occurrence of most types of image compositing operations. In response to these two studies, the thesis presents an alternative approach to digital image forensics, based on automated plan generation. By treating the image inspection process as a plan comprised of different steps, it proposes an architecture that is able to guide an analyst choosing the next best step for inspecting an image. The generated plans are flexible, adapting on the fly to the observed results. The plans are based on a formal modelling of current forensics knowledge and techniques, so that they can be translated in steps to be executed. The thesis then shows that the limits of such an approach lie in the difficulty to validate results, which is a consequence of the setup of forensics problems: they are problems of distributed trust among parties with limited information. This scenario is analyzed from different perspectives in search for the practical limits of Digital Image Forensics as a whole. The results of such an analysis suggest that the field is lacking in providing practical and accessible solutions to society due to limited engagement in multidisciplinary research rather than due to limited technical proficiency. The thesis then discusses how paradoxes from philosophy, mathematics, and epistemology arise naturally in both real forensics scenarios, and in the theoretical foundations of the field. Digital Image Forensics ultimately deals with human communication and, as such, it is subject to all its complexities. Finally, it is argued that the path for providing useful solutions for society requires a collective engagement from different disciplines. It is the responsibility of the forensics community to develop a common, accessible epistemological framework for this collective enterprise.

Keywords: Digital Image Forensics. Image Composition. Automated Planning. Human-based computation.

RESUMO

Além da Análise, Forense, e de Imagens: Em Busca da Regulamentação de Confiança em Comunicação Multi-Mídia

Esta tese discute o papel da Análise Forense de Imagens como reguladora de mídia digital na sociedade. Isto inclui um estudo com mais de 400 indivíduos para determinar suas capacidades de detectar edições em imagens. Os resultados desse experimento indicam que humanos são facilmente enganados por imagens digitais, tendo dificuldades em diferenciar entre imagens pristinas e editadas. A tese então analisa a efetividade do arsenal de análise forense de imagens contra o estado-da-arte de composição de imagens. Através da análise de padrões fundamentais de imagens, as técnicas forenses são capazes de detectar a presença da maioria das operações de composição testadas. A tese então apresenta uma abordagem alternativa para análise forense de imagens, baseada na geração automática de planos. Ao tratar o processo de inspeção de uma imagem como um plano composto de múltiplos passos, propusemos uma arquitetura que é capaz de indicar os passos necessários para analisar uma imagem. Os planos são baseados em uma modelagem formal do conhecimento e técnicas forenses, de modo que possam ser traduzidos em passos a serem executados. A tese então demonstra que os limites de tal abordagem dependem da dificuldade de validar tal solução. Isso é uma consequência da natureza dos problemas de análise forense de imagens: essencialmente, são problemas de confiança distribuída entre indivíduos com acesso limitado à informação. Essa configuração é analisada de diferentes perspectivas em busca dos limites práticos para a análise forense de imagens digitais. Os resultados dessa análise sugerem que a área falha em produzir soluções acessíveis para a sociedade não por limitações técnicas, mas pela falta de um engajamento multi-disciplinar. A tese então discute como paradoxos filosóficos surgem naturalmente em cenários de análise forense de imagens. A análise forense de imagens digitais lida, essencialmente, com comunicação humana e, como tal, está sujeita a todas suas complexidades. Finalmente, é argumentado que o caminho para construir soluções úteis para a sociedade requer um esforço coletivo de diferentes disciplinas do conhecimento. É responsabilidade da comunidade forense desenvolver uma teoria epistemológica comum e acessível para este projeto coletivo.

Palavras-chave: Análise forense de imagens digitais, composição de imagens, planejamento automático, computação baseada em humanos.

LIST OF ABBREVIATIONS AND ACRONYMS

DIF	Digital Image Forensics
ELA	Error Level Analysis
CFA	Color Filter Array
PRNU	Photo Response Non-Uniformity
SB	Sleeping Beauty
JTB	Justified True Belief

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Digital technology has evolved to a point where it is possible to composite and edit images in ways that are virtually undetectable to the naked eye. Moreover, computer graphics techniques can generate realistic pictures that can be hard to distinguish from actual photographs. In such a scenario, how can we tell apart real images from fake ones? In the age of smartphones, online media, and editing software, this is a crucial question.

The subject of Digital Image Forensics (DIF) hints to be in the realm of the technical, computational. Most of its challenges, however, did not stem from technical limitations, inefficient algorithms, or NP-problems. What does it mean for an image to be real or fake? As we demonstrate in this thesis, humans are not able to tell real from fake images through visual inspection only, and there are no clear, technical definitions for real and fake images. The field of law, which is one of the most direct applications of DIF, is equally unable to provide a satisfying answer (PARRY, 2009). The use of digital images in law varies greatly from country to country, or even from case to case. Are fake images the opposite of real images, in the same sense that false opposes true? Are fake images the ones that lack the property of "realness", or rather the opposite, real images are the ones that lack "falseness"? This line of questioning eventually lead us to philosophy and epistemology.

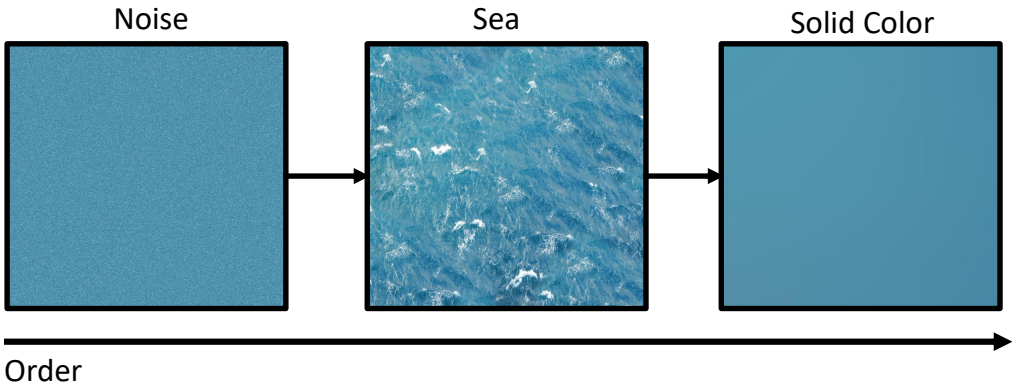
Some of the topics discussed in this dissertation are long-known problems from different fields, and do not have clear solutions. One contribution of This thesis recognizes these problems within DIF, and relates them to their counterparts, be them philosophical, judicial, or computational.

1.1 Images as Information Currency

Many models in Communication Theory can be used to understand the role of digital images in our communication, such as Shannon and Weaver (SHANNON, 2001), or Schramm (SCHRAMM, 1954). In such models, there is generally a distinction between the message, the content, and the channel of communication. Communication happens between two parties A and B through the process of exchanging messages. For explanatory purposes, let us draw on these concepts, and explore the idea of images as *information currency*. Currently in our society, images are a sort of information currency in a similar way money is a *financial currency*.

Images allow the storage of information and its exchange. We can separate the "content", *i.e.*, the visual information, from the storage itself, but only to some extent. If we think about the image format (*e.g.*, jpeg, .png, etc.), resolution, or color depth, these are all characteristics of the message container. Digital images are like a big mosaic, a matrix of tiny elements (pixels) where we can arrange small colored pieces to express a message. But the separation between container and content is not so clear in reality. In Figure 1.1, the mosaic is limited by the same blue-colored pixels, which are only ordered differently. Is the the content the configuration of colored pixels, or the symbolic representation it is purposed to evoke?

Figure 1.1: Transformation of an image's subject as its colors become increasingly ordered. Both a noise image or a solid color seem purely synthetic, or abstract, in contrast to a picture of the sea. However, when viewing each of them side by side, this distinction becomes subjective. The second image is obtained by re-ordering all pixels from the first one. This process is far different from the physical interactions between water, light, and our visual system, yet it can be used to evoke the same subject "sea". This image is then further ordered, averaging its colors to obtain a single, solid color. At which point in the transition between the first and second images the subject of "sea" emerged, and at which point it ceased to be, fading into a solid color?



Money allows people to store their work or value, and exchange between one another. In this sense, it is **not only** paper, coin, or digital money, but the whole system in place, along with the trust of the participants. This has been widely understood in economics since the Roman empire, and its effects can be seen, for instance, in inflation. The practical aspects of this currency has direct impact in our economy and in our society. For a long time the value of currency has been tied to its materials: a one euro gold coin would have one euro worth of gold in its coinage, for example. This means that if the value of the material went up or down, the weight of the money someone had was more

significant than the units. This of course allowed many types of frauds: introducing fake coins into the economy, forging fake coins with cheaper metals, scraping coins for material, etc. Society had to adapt to these issues to maintain trust in coins, and many clever solutions were found to the problem, like adding markings to the coins. In those days, one should be able to tell real from fake coinage analyzing various aspects such as weight, shape, color and seal. People who did not possess such appraisal skills were at risk of being cheated.

The spread of digital images happened much faster than society was able to develop the same appraisal skills that we have for money. It is a lot easier to take photos and edit images than it is to analyze them. This leaves our "information economy" in a delicate position. It has been shown that fabricated visual evidence can be used to convince witnesses (WADE; GREEN; NASH, 2010), and that fake news (ALLCOTT; GENTZKOW, 2017) has a great potential for spreading and affecting our judgment. In the same way that the material aspect of money had an impact on its usage, we are still understanding how the "material aspects" of digital technology (*e.g.*, the image container) affect the economy (GOLDFARB; TUCKER, 2017), and society (SIAPERA, 2017) as a whole.

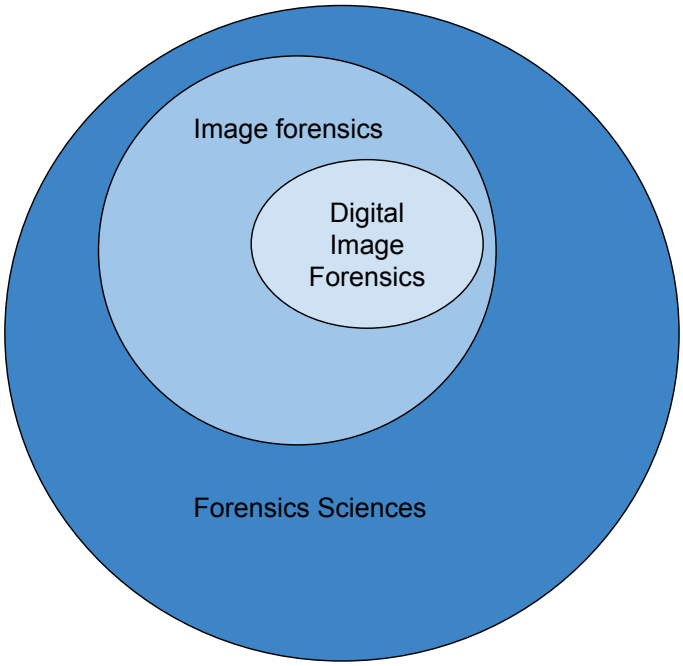
Digital Image Forensics has been charged with the responsibility to regulate this new "information currency". This thesis evaluates the current efforts, progress and challenges of DIF in this complex environment. It expands the discussion beyond the technical realm into the obligations of DIF, putting in perspective what one *can*, and what one *should* be demanding from it.

1.2 What is Digital Image Forensics

The term *digital image forensics* has vague definitions. It stands in the crossroads of many concepts, and the relations among them are not completely clear. It is possible to see it as a specialization of forensics sciences (Figure 1.2), which would imply a strict connection with criminal investigation and its applications in courts of law. Research on digital image forensics reveal a complex network of specialized knowledge from several fields (Figure 1.3). There are arguably more fields that could be considered in this network, but the idea of exhaustively listing fields within DIF is discussed further in the text.

One of the main issues discussed in this thesis is the gap between the theoretical and the practical aspects in DIF. It seems rather paradoxical that it is easier to explain

Figure 1.2: Venn diagram showing Digital Image Forensics as a sub-field of the Forensics Sciences.



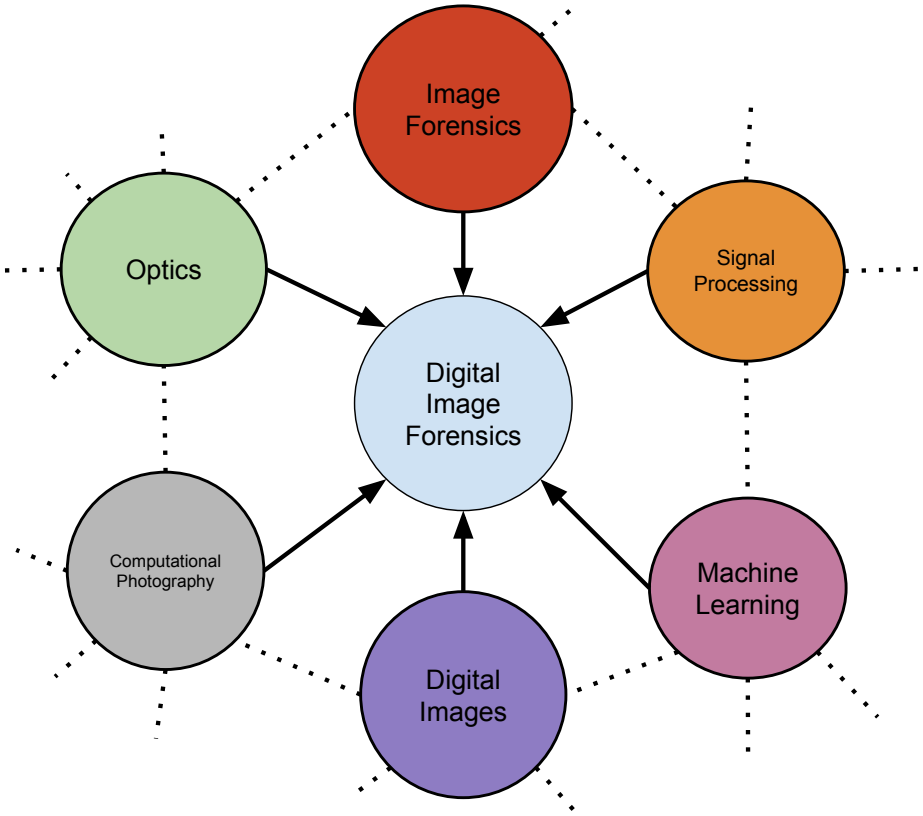
what DIF is (or should be) to someone the least (s)he is associated to one of its related fields. An elegant summary that seems to get the message across for most people is: "it is sort of CSI[1] with images". The mental picture formed by this analogy is more palpable than any detailed description found in the literature.

The effort to define DIF has been constant during my research, and the informal explanation provided in the previous paragraph is just one of the possible answers. Several disciplines *technically outside* DIF turn out to be essential to it. Studies on human perception are an example, and are discussed in Section 2.4. The importance of adequately defining a field, its goals, its scope, is to ensure that research effort is being properly directed. A simple and effective definition then follows: DIF should be *the field concerned with the regulation and understanding of digital images as information currency*. It can be argued that such definition is beyond DIF, and that such field, should it exist or be created, could have a more general, descriptive title.

Figure 1.4 shows the three fundamental components of DIF: *Digital*, *Image*, and

[1]Crime Scene Investigation, which refers to the practice of investigating a crime scene, but also to the famous television series with the same name.

Figure 1.3: A representation of Digital Image Forensics considering possible related fields and their connections. DIF draws immediately from many different fields (black arrows), some of which are related between themselves, sharing theoretical background or methods (dotted lines).



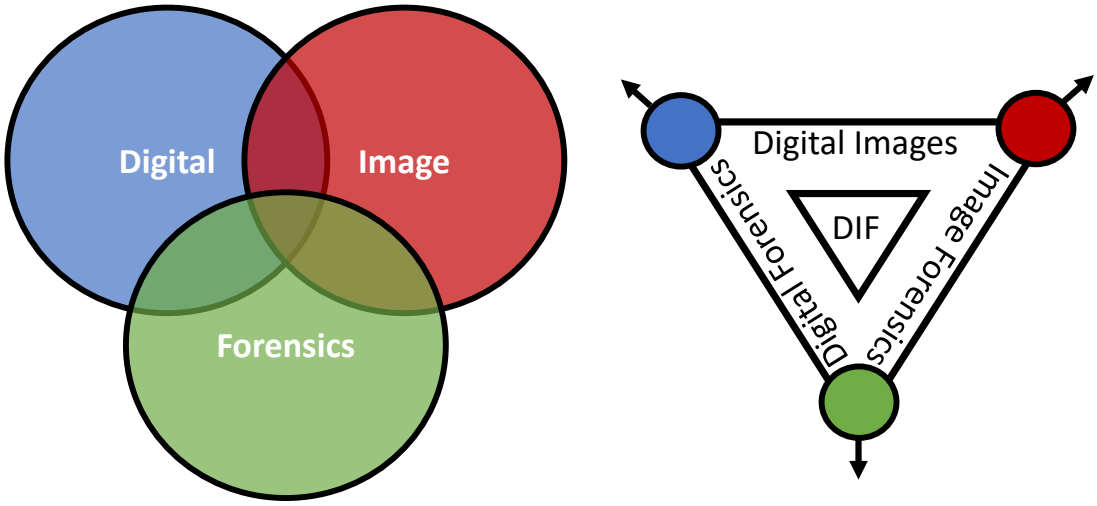
Forensics. The study of *Image* meets with the study of *Digital*, together with the study of *Forensics*. There is no limit to how much knowledge about images one needs for DIF, neither about digital, or forensics for that matter. Therefore, the actual scope of DIF is the union between these three sets, pinning the three fields of study together. In some sense, it represents the powerset of {Digital, Image, Forensics}, where all the combinations between the elements are meaningful subjects. The thesis discusses the role of each individual member (*i.e.*, Digital Image Forensics, Digital Images, Digital Forensics, Image Forensics, Digital, Image, and Forensics) in the regulation and understanding of information currency.

1.3 Thesis Statement

The central idea of this thesis can be summarized in the following statement:

Humans can be easily fooled by digital images, yet such images are essential for

Figure 1.4: A representation of Digital Image Forensics



communication (Chapter 2). The current array of technical solutions to assist in the appraisal of digital images is vast, but has limited utility in real scenarios (Chapters 3 and 4). The nature of this limitation is beyond technical, in the same way the applications of digital image forensics (DIF) are beyond technical, covering social, judicial, and economical aspects, among others (Chapter 5). To properly bridge the technical and non-technical aspects of DIF one will need a common theoretical background. In that regard, most challenges are of philosophical nature, and could be tackled through the advancement of an epistemological theory for DIF (Chapter 6).

1.4 Contributions

The **contributions** of this thesis include:

- Experimental evidence that humans have difficulty to detect forgery in digital images, even in a context where they have been explicitly told to look for it (Chapter 2);

- A dataset of subjects' answers for real and forged images, with 17,208 answers over 177 images, and 8,160 image markings indicating what subjects considered to be forgeries;

- A Forgery Detection scale for forensics assessments that classifies the output of forensics techniques according to the type of uncovered information (Section 3.1);

- A review of the state-of-the-art in Image Composition from a forensics point of view, organizing techniques by their forgery potential, and discussing their strengths and weaknesses against detection (Section 3.3);

- A qualitative analysis of several composition techniques to an uncontrolled image set, showing promising interactions with traces for exploration (Section 3.4.1);

- A quantitative analysis on a fully controlled image set, providing an in-depth analysis of the behavior of double JPEG compression on different composition techniques (Section 3.4.2);

- A novel approach to DIF based on automated plan generation (Chapter 4);

- An argumentative analysis on the technical limitations of DIF to solve practical problems (Chapter 5);

- A discussion about the epistemology of Digital Image Forensics, and the challenges for producing scientific knowledge in such multidisciplinary field (Chapter 6).

1.5 Thesis Structure

The thesis is organized according to the structure presented in the Thesis Statement. Each Chapter is self-contained, associated to one of the ideas presented, and building on the ongoing narrative as follows:

In Chapter 2, we argue that humans are easily fooled by digital images. It describes a study providing experimental evidence about humans not being suited for distinguishing real from fake images. The chapter also provides lengthy validation of the collected data (Section 2.3), and shows that different behavioral features correlate with subjects' performance.

Chapter 3 analyzes the array of technical solutions in DIF. First, it introduces a scale to analyze the output of forensics techniques. The fields of DIF and image composition are surveyed respectively on Sections 3.2, and 3.3. From the meta-analysis of both fields, it re-organizes the literature on image composition, proposing a classification based on potential for forgery. Section 3.4 presents a series of experiments in which many image composition techniques have been tested against forensics ones. The chapter shows that the field of DIF is very well-served technically to detect most existing types of forgery.

Chapter 4 proposes a novel approach for forensics analysis through the use of automated plan generation. Section 4.1 introduces the conceptual basis for automated planning, and its relation to DIF. Our strategies to implement this are discussed in Sections 4.1.1 through and 4.1.3. The architecture of our solution is explained in Section 4.2, and the contributions of our prototype are detailed in Section 4.3. Finally, we present the challenges to validate such an approach, and the implications of this for DIF (Sections 4.3.1, and 4.3.2).

Chapter 5 builds on previous discussion to show that the applicability of DIF is limited by communication, rather than technology. We construct real-world cases for the use of DIF, and dissect them from a practical perspective (Sections 5.1 to 5.3). We use a thought experiment to challenge the technical aspects of DIF (Section 5.4), and its different uses of probability (Section 5.6.1).

Chapter 6 provides an analysis on the epistemologic foundations for DIF knowledge. It presents the classical interpretation of knowledge as justified true belief, the attack by Gettier cases, analyzes the nature of evidence in DIF. The chapter then discusses the difficulty for using mathematical models and machine learning as justification (Section 6.1.1). It shows that important challenges to model knowledge for DIF arise from classical problems in philosophy (Sections 6.3, and 6.4).

Chapter 7 closes the dissertation by revisiting the power set of {Digital, Image, Forensics}, showing that it presents a meaningful depiction of the responsibilities of DIF. The different types of problems discussed in this thesis are a result of this accountability. We outline a new perspective for the field that reaches out to different disciplines in pursuit of fulfilling its much needed role in society.

2 HUMANS ARE EASILY FOOLED BY DIGITAL IMAGES

This Thesis begins by justifying the need for the field of Digital Image Forensics, which provides technical solutions to the appraisal of images. The idea that humans are not suited to assess the authenticity of images without the aid of tools has been widely accepted in the forensics community ((PIVA, 2013), (ROCHA et al., 2011), and (FONTANI et al., 2013)). However, studies on human perception of digital images have focused on very specific aspects of vision such as color ((XUE et al., 2012b), and (LALONDE; EFROS, 2007)), lighting ((OSTROVSKY; CAVANAGH; SINHA, 2005), and (RADEMACHER et al., 2001)), geometry ((FARID; BRAVO, 2010), and (VISHWANATH; GIRSHICK; BANKS, 2005)), and face recognition ((SINHA et al., 2006), (GAUTHIER et al., 2000), and (FAN et al., 2014)). However, before ours, no extensive study has been performed to evaluate one's ability to detect editing in digital images.

This chapter, based on our paper entitled "Humans Are Easily Fooled By Digital Images" (SCHETINGER et al., 2017b), provides evidence that supports the hypothesis that humans are not good at identifying image forgeries. For this, we performed an experiment with approximately 400 subjects. The experiment was specifically designed to avoid guessing, requiring evidence to support the subjects' answers. The results show that only 58% of the images were correctly classified as either pristine or edited, and only 46% of the edited images were identified as such, *i.e.*, *more than half of all edited images were unnoticed.* This performance is superior to random guessing, as we show in our validation, but lower than most computational forensics techniques. To make the experiment as relevant as possible for the forensics community, we used images from known public forensics datasets.

Our study differs from previous ones because it requires evidence whenever the subject believes the image has been altered (*i.e.*, (s)he should point in the suspected image region). This allows us to discard lucky guesses, and also provides insights on what subjects perceive as being suspicious in an image. To be able to gather a large amount of data, we performed an on-line experiment. Due to the uncontrolled nature of on-line tests, we apply a series of validation checks to the collected data, and discard answers containing inconsistencies. We show that our results are statistically significant.

The **contributions** of this chapter include:

- Experimental evidence that humans have difficulty to detect forgery in digital images, even in a context where they have been explicitly told to look for it (Section

2.2.1);

- A dataset of subjects' answers for real and forged images[1], with 17,208 answers over 177 images, and 8,160 image markings indicating what subjects considered to be forgeries;

- Evidence that age, experience with digital images, and answering behavior of a subject, such as timing and confidence, affect one's performance when looking for forgeries in images (Section 2.2.2);

In addition to these contributions, we discuss how different image features may correlate with certain types of answers (Section 2.2.3), and with subjects' perception of the test difficulty (Section 2.2.4). The lengthy validation process for the experimental process, which was published as supplemental material is reproduced in Section 2.3.

2.1 The User Study

The goal of our study was to assess how hard it is for an average individual to determine if an image has been modified. For this, we gathered input from a large group of subjects over a large image database. Subjects are shown one image at a time and asked to provide a binary *yes/no* answer to the following question: "Is there any kind of forgery in this image?". For simplicity, we call an *authentic* image (also referred to as *pristine* or *original*, in the forensics literature) as a T (*true*) image. Likewise, we will call a *modified* image (also denoted *forged*, *tampered*, *fake* or *edited*) as an F (*false*) image. If a subject answers *yes*, (s)he means that the image is *false*, and we call this an $\mathbb{F}$ answer, as opposed to a $\mathbb{T}$ (*true*) answer. In this case, the subject is asked to *provide evidence that the image has been altered*. Such evidence is given by pointing to an image region that indicates it has been altered. Different forms of evidence are considered valid, such as the altered region itself, its close surroundings, or even irregular shadows left by the forgery. *For* F *images, an answer is considered correct only if valid evidence has been provided*.

Considering all the different answer combinations, there are five possible outcomes: the image can be either T or F, the subject answer can be either $\mathbb{T}$ or $\mathbb{F}$, and if the subject answers $\mathbb{F}$, (s)he can provide either valid or invalid evidence (Table 2.1).

For consistency with the forensics literature, *we treat the subjects' answers as a binary classification problem of identifying* F *images*. Thus, a *true positive* consists of

[1]The dataset will be made available upon paper acceptance.

Table 2.1: Different answer classes for the subject study, in the notation “Image Type:Answer Type".

Class	Meaning	Answer	Type
T:$\mathbb{T}$	The image is T and the subject provided a $\mathbb{T}$ answer.	Correct	True Negative
F:$\mathbb{F}$v	The image is F, the subject provided a $\mathbb{F}$ answer and valid evidence.	Correct	True Positive
F:$\mathbb{T}$	The image is F and the subject provided a $\mathbb{T}$ answer	Incorrect	False Negative
F:$\mathbb{F}$i	The image is F, the subject provided a $\mathbb{F}$ answer and invalid evidence.	Incorrect	False Negative
T:$\mathbb{F}$	The image is T and the subject an $\mathbb{F}$ answer.	Incorrect	False Positive

answering $\mathbb{F}$ and providing valid evidence to an F image (F:$\mathbb{F}$v). A true negative, then, consists of answering $\mathbb{T}$ to a T image (T:$\mathbb{T}$). A *false positive* consists of answering $\mathbb{F}$ to a T image (T:$\mathbb{F}$). Finally, a *false negative* consists of either answering $\mathbb{T}$ to an F image (F:$\mathbb{T}$), or answering $\mathbb{F}$ to an F image, but failing to provide valid evidence (F:$\mathbb{F}$i) (see Table 2.1).

2.1.1 The User Study

For our on-line experiment, subjects were asked to register, providing background information such as age, education, and experience level with digital images. Once registered, subjects could log in at any time to analyze and classify images, being able to interrupt and resume the classification at their convenience. The answering form consisted of a simple web page, as depicted in Figure 2.1. After observing an image for at least 20 seconds, the subject could ask for a hint, which consists of removing a rectangular region corresponding to half of the image area not containing any editing. In the case of a T image, a randomly positioned rectangular area is used on one of the sides or the center of the image. The total area removed is always half of the image, and the image always remains contiguous.

Each F image from the test database has an associated binary mask covering a region of the image considered as the location of valid evidence for the forgery. Such mask is called the *evidence evaluation mask* and is used for two purposes: to evaluate if the evidence provided by the subject is valid; and to determine what parts of the image can be discarded to provide a hint to the subject. The evidence evaluation masks have been created using, as source, the ground truth binary masks of pixels changed in the doctoring process of each image. To cover different kinds of evidence, the masks were edited by hand increasing the valid area. Thus, for instance, the ground truth mask on Fig. 2.1b only depicts the trophy added to the image. The evidence evaluation mask (Fig. 2.1c), on the other hand, contains a larger area. In this case, both the lack of shadows on Fig. 2.1b

Figure 2.1: Interface for the on-line user study. The currently evaluated image is shown at the center. Radio buttons register the subject's answer (Yes/No) and confidence level (Low/Medium/High) for the answer. A menu (top right) displays the subject's progress, and provides access to other options.

and the region around the trophy edges are also considered valid evidence. Each F image was carefully evaluated to construct its evidence evaluation mask, as this is subjective and context dependent.

The data collected in the user study consists of: (1) subject answer (Yes / No); (2) evidence in the form of a click (when answered "Yes"); (3) confidence level in the answer (Low / Medium / High); (4) did the subject request a hint? (Yes/No); (5) subject observation time before asking for a hint; (6) subject observation time after asking for a hint; and (6) did the subject observed the image in its original resolution? (Yes / No).

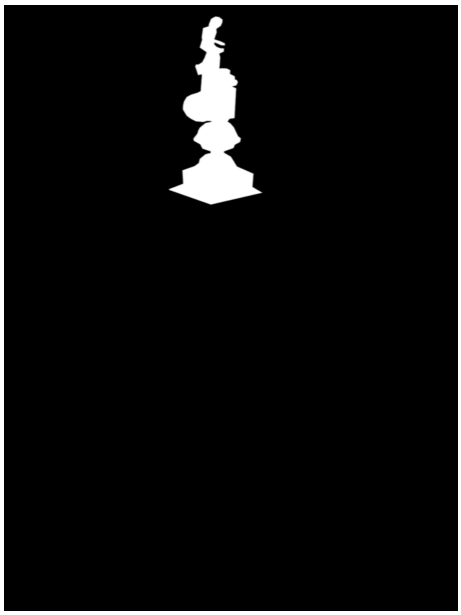
2.1.2 Image Database

Our image database consists of 177 images, divided into 80 true images (45%) and 97 false images (55%). The false images consist of 20 erasing images, 35 copy-and-paste images, and 42 splicing images. An *erasing forgery* consists of using brushes, blurring, or even copying some small patches to hide some portion of the original image. A *copy-paste* forgery consists of copying from, and pasting on, the same image, some region or object with or without transformations such as scaling and rotation. Finally, a *splicing*

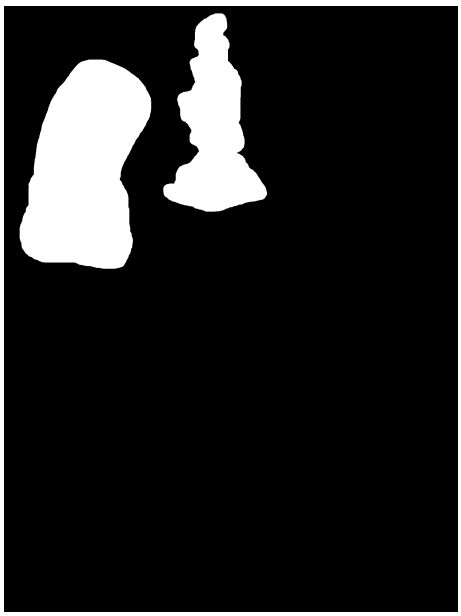
Figure 2.2: Demonstration of the hint provided to the subject for an image and the masks used in the process. In this case, the trophy was added to the pedestal. From left to right, the second picture is the ground truth edition mask, depicting the exact pixels that were changed in the manipulation. The third picture is a manual mask drawn over the ground truth mask that delimits all the regions in the image that can be considered valid evidence. In this case, the trophy, its surroundings and the shadow area. The figure on the right depicts the activated hint, blackening out a part of the image irrelevant to the process of finding manipulation.

(a) Image without hint.

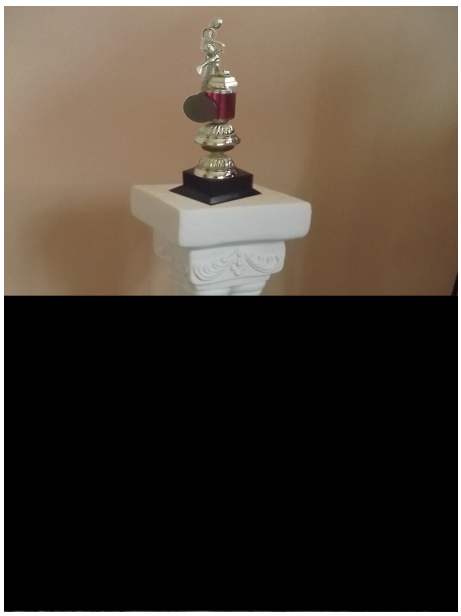
(b) Ground truth edition mask.



(c) Evidence evaluation mask.



(d) Image with hint.



forgery consists of copying a region from an image and pasting over another image, also with the possibility of transformations (see Figure 2.3).

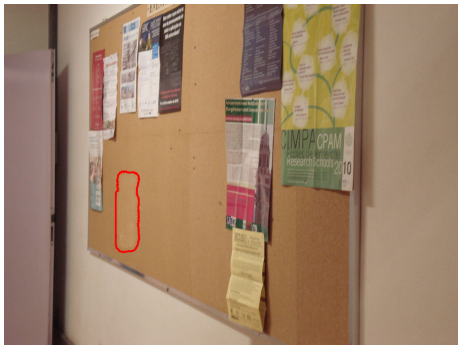
The images used in our user study have been handpicked from three public foren-

sics image databases: the *forensics challenge database* (MINUTES..., 1948)(IEEE, 2013), the *splicing database* provided by Carvalho et al. (CARVALHO et al., 2013), and Cozzolino et al. (COZZOLINO; POGGI; VERDOLIVA, 2014) *copy-and-paste* database. The total image count adding all databases is around 6,000 images, with a great majority consisting of true images and false images with splicing operations.

Figure 2.3: Examples of different image types present in the assembled database. The edited area of F images is outlined in red.

(a) T image.

(b) F image with an erasing forgery. Papers have been erased from the board.



(c) F image with a copy-paste forgery. One of the cisterns has been duplicated.

(d) F image with a splicing forgery. The man's face has been spliced from a different image.

The motivation for using known and public forensics databases was to avoid the bias of performing a test with self-made forgeries, and to use images that are commonly analyzed by forensics techniques for comparison. To reach the final number of 177 images, an *iterative process* was used to select images from the original pool into subsequently smaller pools. The following selection criteria were used:

- *Multiple image types*: our image database should include several examples of true images, as well as erasing, copy-and-paste, and splicing forgeries. This allows us to compare the results between different image types. Figure 2.3 depicts examples of each type of evaluated forgery;

- *Image context:* the images were classified according to whether they depicted nature, people, buildings, landscapes, and if they were taken indoors or outdoors. Using these criteria, we intended to achieve two main objectives: generalize our results in a way to cover a wide set of test scenarios, and keep subjects interested during the experiment.

Preference was given towards visually easy-to-detect forgeries, and scenes depicting people. Images with edited areas larger than half of the image size, or with multiple edited regions were considered inappropriate for the test, because they would conflict with the designed hint system (*i.e.*, removing a rectangular area covering 50% of the image pixels).

The manual image-selection process was important because the used databases were designed mostly for non-assisted forensics techniques, resulting in a large number of forgeries that are not meaningful for humans. An example of this would be copying and pasting two regions with the same color one over another. It might be trivial for a forensic technique that evaluates PRNU[2] (CHIERCHIA et al., 2014) or compression artifacts (BIANCHI; PIVA, 2012a) to identify this type of forgery, but it makes little sense to ask a human subject to do so. Only 20 erasing cases were selected because of the lack of adequate images in the databases.

To evaluate the diversity of content of our final database, we made two additional validation processes. First, we asked forensics experts to analyze all F images and rate the difficulty in detecting the forgeries. Secondly, we tagged all images according to scene elements, such as indoor/outdoor, natural/artificial light, contains people, contains buildings, contains animals, contains household objects, etc. According to the forensics experts, *the majority of the* F *images contain modifications that are easy to detect*. Thus, *any possible bias caused by image difficulty should influence user performance positively.* Full details about the used methodology and the results are available in Section 2.3.

2.1.3 User Motivation and Usability

The task of analyzing 177 images can quickly become boring and underwhelming for human subjects. This is a serious issue that guided the design process of our user study. Two main approaches were used to tackle it: providing motivation for subjects

[2]Photo Response Non-Uniformity (PRNU), a form of noise pattern for photo sensors.

to go on, and *ensuring that each answer could be treated equally, independently of the subject and how many images (s)he analyzed in total.*

To motivate subjects to finish the study, practices from *serious games* (RITTERFELD; CODY; VORDERER, 2009) were used: the answering process was divided into 10 different levels, and at each provided answer the subject gathered experience points to progress to a next level. The participants kept track of this information through the interface. Upon finishing a level, the subject's performance was calculated for the answered images on that level, with statistics such as hit rate, average time, confidence, and hint usage. Note that these statistics are based on all images analyzed on the finished level. Thus, *it is not possible for a participant to determine exactly which images were answered correctly.* This information is saved as part of the subjects' profile and can be reviewed by us at any time, summarizing their performance for each finished level. This is an important functionality, because subjects need constant feedback. However, providing feedback on each answer would compromise the study results.

Upon finishing all levels and completing the test, a subject earned the right to appear on the high scores page, where one's overall statistics are displayed. This was included after suggestion from participants in a pilot test. They argued that they felt motivated by the competition aspect. It is important to note that the level progression system does not affect, in any way, the order of images, nor it presents any change in difficulty. It serves merely as a motivational and progress keeping mechanism.

The order in which the images appear for each subject is semi-random, based on an algorithm that prioritizes images that so far have received less answers. The algorithm randomly selects an image from the pool of least answered ones. Reloading the page or coming back another time evokes this process again, changing the image. Once an image has been answered by the subject, it cannot be selected again for him. This guarantees that all images have a similar amount of answers, and that the collected data is as homogenous as possible regarding to answer distribution.

The user interface was designed to be as user-friendly and consistent as possible. It was tested on the most common browser resolutions, displaying the image in fixed resolution of 1024×768 for all subjects. The fixed resolution is important for consistency, but *all images can be inspected in their original resolutions* by selecting the link under it.

Since the study is based on visual aspects and can be exhaustive for subjects, special effort was made to minimize stress and confusion. The interface design took several iterative cycles, and considered aspects such as colors, button size and placement,

text, menu size and placement, and tooltips. All textual information provided is bilingual, both in English and Brazilian Portuguese.

To ensure that subjects were not influenced by our feedback, or improved their results with experience during the test, we compared their performance along different levels. *We found that the distribution of answer classes remained consistent for all levels, both individually and globally.* Further discussion and validation of our approach is provided in in Section 2.3.

2.2 Results and Discussion

This section analyzes the collected data. Since we extend a traditional binary classification problem, we use the term *performance* to denote the same concept as traditional accuracy (*i.e.*, the number of correct answers over the total number of answers). This is done to avoid confusion with answers that were classified correctly but answered wrong due to invalid evidence. The subjects' answers are classified according to the criteria shown in Table. 2.1.

To determine if two features are correlated, we estimate the Pearson coefficient ρ, as well as the $p-$value, which can be defined as the *smallest value* for the significance level α of the statistical test. If the p-value is smaller than α, the correlation between the features is considered significant (TRIVEDI, 2002).

Some features, such as Confidence and Time, are discussed in both Subsections 2.2.2 and 2.2.3, but under a different point of view. On Subsection 2.2.2 we analyze the subjects, so their features are averaged over all of their answers, obtaining a feature vector for each participant (393 in total). On Subsection 2.2.2 we focus on the images, determining their features averaging all the answers it received from different participants. This results in one feature vector for each image (177 in total). In both cases all 17,208 individual answers are used, but they aggregated in different ways.

2.2.1 Overview of Results

We collected a total of 17,208 answers from 393 different subjects, after discarding invalid entries using two validation tests (explained in detail in the supplementary material). The first validation test simulates a subject randomly answering *Yes/No*, as well as

Table 2.2: Distribution of the 17,208 valid answers according to various criteria.

	Total	Proportion
T Images Answered	7,791	0.452
F Images Answered	9,471	0.548
$\mathbb{T}$ Answers	9,048	0.525
$\mathbb{F}$ Answers	8,160	0.475
Answered Correctly	9,899	0.576
Answered Wrong	7,309	0.424
Erasing Images Answered	1,942	0.113
Copy-Paste Images Answered	3,392	0.197
Splicing Images Answered	4,083	0.237
T:$\mathbb{T}$ Answers	5,520	0.320
F:$\mathbb{F}$v Answers	4,379	0.254
T:$\mathbb{F}$ Answers	2,271	0.132
F:$\mathbb{F}$i Answers	1,510	0.087
F:$\mathbb{T}$ Answers	3,528	0.205
Requested Hints	3,091	0.179
Number of Full-Res Visualizations	1,744	0.101

providing random evidence. This test checks whether our methodology is *robust against random guesses*. The second validation test consists of resampling subsets of the total data using only balanced amounts of answers from each subject. This is used to evaluate if the use of an unbalanced number of answers among subjects would bias the results.

Tables 2.2 summarizes the results of our experiment, including the distributions of image classes, types of valid answers, and numbers of requested hints and images visualized at full resolution. Figure 2.4 shows the overall distribution of classes between all answers, and how they compare to our validation tests. In this stacked bar graph each column represents a different scenario (simulation, resampling, and original data), and all the bars in each column have values adding to 1. The blue bars represent the correct answer classes (T:$\mathbb{T}$ and F:$\mathbb{F}$v), and the red bars the wrong answer classes. In this way, in each scenario, one can simply measure the performance as the height of the two stacked blue bars.

This figure shows that *the distribution of answers obtained in the experiment (Original) differs greatly from random guesses (Simulation)*. Also, *the balance of the number of answers per subject had little impact on the distribution of answers (Resampling) when compared to the original (Original)*. Both the random simulation and resampling data shown in Figure 2.4 are the average of 1,000 generated answering distributions considering 17,700 and 10,051 images, for the cases of random simulation and resampling, respectively.

Figure 2.4: Comparison of the answer distribution among random simulations, balanced resamplings, and the original dataset. The random simulations reproduce a subject guessing image types and evidence. The balanced resampling equalizes the number of different forgery types answered by each subject. Both the simulation and resampling data displayed are the average of 1,000 generated datasets. Notice how the original (full) dataset greatly deviates from the random simulation, while having almost the same class distribution as the balanced resample.

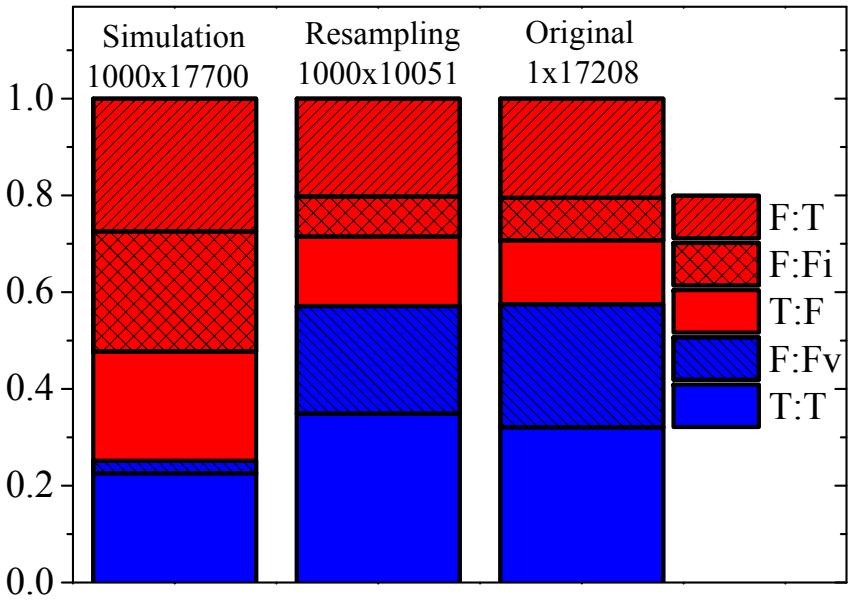


Table 2.3 presents the classification statistics for our experiment. *Performance* is the amount of correct answers over total answers. *Precision* is the amount of true positive answers over all true positive and false positive answers. *Specificity* is the amount of true negative answers over all negative answers. *Sensivity* is the amount of true positive answers over all positive answers, and *F1 Score* is the harmonic mean of precision and sensivity.

An inspection of Table 2.3 reviews that the subjects did not find anything suspicious in the majority of the edited images. A sensitivity value of 0.465 means that only 46.5% of the answers for $\mathbb{F}$ images were correct. A specificity value bigger than the sensitivity value indicates that the subjects provided $\mathbb{T}$ answer more often. This is also corroborated by the data on Table 2.2: while the majority of tested images are $\mathbb{F}$ rather than $\mathbb{T}$ (9,471 against 7,791), subjects provided more $\mathbb{T}$ answers (9,048 against 8,160). This is understandable, as providing a $\mathbb{T}$ answer requires no additional evidence. When unsure, subjects seem more prone to answer $\mathbb{T}$.

Table 2.3: Classification statistics for the subjects answers. Table 2.2 contains the values used in the formulas.

Statistic	Formula	Value
Performance	$\frac{\mathrm{F}:\mathbb{F}v+\mathrm{T}:\mathbb{T}}{\mathrm{F}:\mathbb{F}v+\mathrm{F}:\mathbb{F}i+\mathrm{T}:\mathbb{T}+\mathrm{T}:\mathbb{F}+\mathrm{F}:\mathbb{T}}$	0.575
Precision	$\frac{\mathrm{F}:\mathbb{F}v}{\mathrm{F}:\mathbb{F}v+\mathrm{T}:\mathbb{F}}$	0.658
Specificity	$\frac{\mathrm{T}:\mathbb{T}}{\mathrm{T}}$	0.708
Sensitivity	$\frac{\mathrm{F}:\mathbb{F}v}{\mathrm{F}}$	0.465
F1 Score	$\frac{2(\mathrm{F}:\mathbb{F}v)}{2(\mathrm{F}:\mathbb{F}v)+\mathrm{T}:\mathbb{F}+\mathrm{F}:\mathbb{F}i+\mathrm{F}:\mathbb{T}}$	0.545

The numbers in Table 2.3 suggest that human subjects are less capable of detecting forgeries in images than existing computational forensics techniques. Considering only works that have images in common with our database, Cozzolino et al. (COZZOLINO; POGGI; VERDOLIVA, 2014) and Chierchia et al. (CHIERCHIA et al., 2014) have consistently achieved F1 Scores over 0.8 for the copy-paste images; Carvalho et al.'s (CARVALHO et al., 2013) best configuration for spliced images reaches up to 0.68 sensitivity and 0.9 specificity. The Ranking of the International Forensics Challenge uses a combination of different metrics, with both image and pixel-wise statistics for classification. Details about the actual metric can be found in (IEEE, 2013). The best contestants achieved a nearly perfect identification rate. Figure 2.5 illustrates our reasoning for these claims. While it is impossible to achieve a quantitative comparison between our subjects and the techniques, we can qualitatively infer subjects have inferior performance.

2.2.2 Subject Background and Behavior

Our user study was designed to maximize the amount of individual answers. Thus, all answers are treated equally, independent of subject. The majority of subjects did not answer more than 40 images, but over 50% of the valid answers were provided by subjects who analyzed more than 100 images. In total, 46 subjects fully completed the test, having analyzed all 177 images.

Table 2.4 shows the distribution of subjects according to their background information (age, education, and experience level with digital images). The majority of answers were provided by adults between 21 and 35 years old, with college education, and an amateur level of experience with digital images. Such profile is arguably *one of the*

best suited demographics to perceive manipulation in digital images, and it is likely to represent an upper bound for the general population in terms of performance.

To verify how subject background correlates with performance, we estimated the Pearson correlation coefficient. The tables with all correlations can be found in our supplemental material. *Age*, *Education* and *Experience* are considered background, and the following features determine subject behavior:

- *Confidence*: the average confidence level considering all subjects and all answers;
- *Time*: average elapsed time before subjects asked for a hint;
- *Time after Hint*: average elapsed time between subjects asked for a hint and provided an answer;
- *Hint Proportion*: the proportion of answered images in which the subject asked for a hint;
- *Full Resolution*: the proportion of answered images in which the subject opened the image in full resolution.

Performance (proportion of $\mathrm{F}{:}\mathbb{F}\mathrm{v}$ and $\mathrm{T}{:}\mathbb{T}$ over all answers) was found to correlate with confidence (0.14), image inspection at full resolution (0.11), age (-0.14), and experience (0.12). This means that subjects with higher performance tend to analyze the image more carefully by opening it in full resolution more often, and are more confident in their answers. Performance decreases with age, and increases with experience with

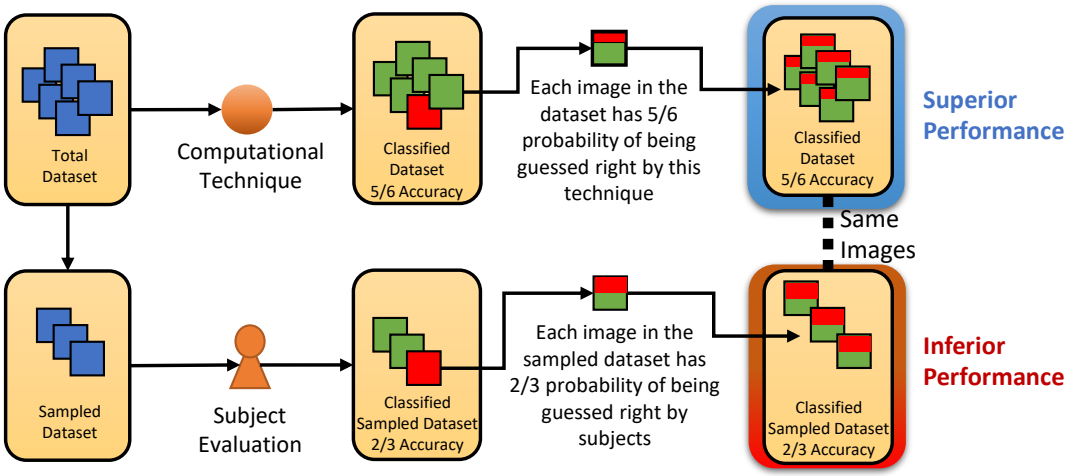
Table 2.4: Overview of the distribution of answers according to subjects's age group, education, and level of experience with digital images.

	Total	Proportion
Age		
Up to 21	513	0.029
21 to 35	14,738	0.856
35 to 50	1,359	0.079
50 and Over	598	0.034
Education		
Highschool	791	0.046
Undergraduate	7,123	0.413
Graduate	9,294	0.540
Experience with digital images		
User	4,621	0.268
Amateur	9,537	0.554
Professional	3,050	0.177

digital images. This might be explained by the younger generations' exposure and familiarity with digital images. Reduction in visual acuity could also play a role with the decrease in performance with age. Education was found to have no significant impact in performance.

Longer analysis time (both before and after asking for a hint) correlate with providing $\mathbb{F}$ answers in detriment of $\mathbb{T}$ answers, regardless of the image type. This suggests that subjects that spend more time analyzing an image are more prone to suspecting it is an F image, even if there is no clear evidence of that. We call this behavior *over-analysis*. A similar observation can be made with respect to the use of hints, with a slightly weaker correlation. If a subject suspects an image is F, the hint seems to increase the suspicion. Such bias might be explained by the nature of the test, since subjects exhaustively scrutinize images for forgeries.

Figure 2.5: Illustration of the performance comparison between computational techniques and our subjects' results. All images used in the test were previously tested by at least one computational technique. Input images from a dataset are represented in blue, and after a testing procedure by a technique (top) or by a subject (bottom), can be rightly classified (green), or wrongly classified (red). Different techniques use different testing methodology and metrics. However, independently of the metric used, they outperform human subjects, which correctly classified the images only approximately half of the time. Note that the numbers used for accuracy in this example ($5/6$ and $2/3$) are merely illustrative.



2.2.3 Image Statistics

We grouped answers according to *image features* and *subject behavior*, and tried to identify correlations between these two classes. To avoid a bias due to the differences in

the image sizes, all images were re-scaled to 1024x768 pixels before feature estimation.

The image features were classified according to the following criteria:

- *Luminance level*: the average relative luminance value of the image using the sRGB color space: $0.2126R + 0.7152G + 0.0722B$ (STOKES et al., 1996);
- *Variability*: average standard deviation of all pixels' relative luminances in a 11x11 neighborhood;
- *Edited area*: ratio between the number of edited pixels over the total number of pixels. It is only computed for F images.

For each image, we also computed subject behavior features. Note that although their names are the same of the subject features explained in Section 2.2.2, they actually have different semantics and are calculated differently. Their meanings are explained below:

- *Confidence level*: average confidence of all answers for the image;
- *Time before hint*: average time before subjects asked for a hint for the image;
- *Time after hint*: average time between subjects asked for a hint and provided an answer for the image;
- *Hint proportion*: percentage of answers for which a hint was requested for the image;
- *Full resolution*: percentage of answers for which the image was observed in full resolution.

To estimate the correlation between image features and answering classes, it is necessary to differentiate between T and F images, because they define mutually exclusive classes. Note that for T images, the proportion of $\mathbb{T}$ answers (*i.e.*, T:$\mathbb{T}$) is the performance. The same holds for F images and $\mathbb{F}$v answers (*i.e.*, F:$\mathbb{F}$v). The data is split between Table 2.5 for T images and Table 2.6 for F images. Since there are only two classes for T images, the correlations between T:$\mathbb{T}$ and T:$\mathbb{F}$ are complementary: they have the same p-value, and ρ values with opposite signs.

For T images, only two significant correlations were found. The more subjects have inspected T images in full resolution, the more they tend to answer that the respective images are F. This seems to contradict the expectation that subjects who analyze images

Table 2.5: Correlation coefficient (ρ) and corresponding p-value for image features and the T:$\mathbb{T}$ class. The class T:$\mathbb{F}$ is complementary. Blue denotes positive correlation with acceptable p-value ($p < 0.05$), and red denotes negative correlation with acceptable p-value. Black values do not satisfy the threshold and the null hypothesis cannot be rejected.

	T:$\mathbb{T}$ ρ	T:$\mathbb{T}$ p
Confidence level	-0.062	0.584
Time before hint	-0.106	0.350
Time after hint	-0.200	0.075
Hint proportion	-0.167	0.138
Full resolution	-0.261	0.020
Luminance level	0.105	0.352
Variability	0.230	0.040

in full resolution tend to give more accurate answers. We also found that inspecting an image at full resolution is associated with lower confidence, higher hint use, and longer time observing the image after asking for a hint.

Variability is positively correlated with the T:$\mathbb{T}$ class, but also correlates with lower confidence level (-0.15), higher hint use (0.17), longer time observing the image after asking for a hint (0.37), and inspection of the image in full resolution (0.17). Our hypothesis is that an image with higher variability is harder to visually inspect, as it contains more details. Thus, subjects might default to a $\mathbb{T}$ answer after not finding anything suspicious.

Figure 2.6 displays four examples of images used in our user study, both T and F with high and low variability. The overlayed versions at the bottom also outline the edited areas and show which image parts the subjects have indicated as evidence of forgery. In Figure 2.5e, a T image with low variability, the subjects promptly suspected of the smoothed sand patterns and the dune being moved by the wind, and it received 50 $\mathbb{F}$ answers (out of 99 total answers). On the other hand, Figure 2.5g, which is also T but with high variability, had an average observation time 6 seconds longer than Figure 2.5e, and only got 16 $\mathbb{F}$ answers (out of 98 total answers).

The *click patterns* indicating the image parts considered suspicious by the subjects also provide valuable insights on the subjects' decision processes. They are represented by heat maps in the overlayed images. The majority of clicks on Figure 2.5a concentrates on a wave of sand (caused by the wind), as it is the most visually striking element of the image. In Figure 2.5d, some subjects have noticed residues from the splicing composition, specially on the back of the bus, and used them as evidence. The shadows of the bus and the biker (which is not forged) caught the attention of several subjects that pointed at it. Finally, the signs and writing on the bus were used as evidence of forgery. It is possible

that subjects suspected that the signs have been altered, and not that the bus itself has been spliced. Since we cannot determine their exact intention, any click on the bus must be considered valid evidence.

Figure 2.6: True and fake images with low and high Variance values, respectively. The orange spots indicate the heatmap of subject clicks on the image, while the red borders on the fake images outline the edited area. The evidence evaluation masks for these images are binary images of the edited area with a small dilation. Here the images are shown side-by-side with their overlayed versions to allow the perception of details.

(a) True image with low Variance.

(b) Fake image with low Variance.

(c) True image with high Variance.

(d) Fake image with high Variance.

(e) Heatmap of clicks over true image with low Variance.

(f) Heatmap of clicks, and modified area highlighted over fake image with low Variance.

(g) Heatmap of clicks over true image with high Variance.

(h) Heatmap of clicks, and modified area highlighted over fake image with high Variance.

For F images there are 97 samples (each F image represents a sample, but over 8,667 answers were used to calculate their features) and three classes: F:$\mathbb{F}$v, F:$\mathbb{F}$i and F:$\mathbb{T}$. Furthermore, the fake images are split into three types: erasing, copy-paste and splicing forgeries; all different in nature. When estimating correlations among multiple such classes simultaneously, it is slightly harder to reject the null hypothesis (i.e., reject the hypothesis that there is no correlation) with the amount of collected evidence. Nevertheless, it is possible to outline several significant correlations, as can be seen in Table 2.6.

The most notable correlations in Table 2.6 are related to confidence level: when subjects recognize something suspicious they are very confident in their answers, but will remain unsure if no clear evidence can be found. The negative correlation between confidence level and the F:$\mathbb{F}$i class is probably caused by subjects actively guessing. Time before hint, time after hint, and hint proportion all have significant correlations, as subjects try to obtain additional clues. Ultimately, they fail to increase performance. This means

Table 2.6: Correlation and respective p-value of image features for fake images and the F:$\mathbb{F}$v,F:$\mathbb{F}$i and F:$\mathbb{T}$ answer classes. Here, blue denotes positive correlation with acceptable p-value ($p < 0.05$), and red denotes negative correlation with acceptable p-value. Black values do not satisfy the threshold and the null hypothesis cannot be rejected.

	F:$\mathbb{F}$v ρ	F:$\mathbb{F}$v p	F:$\mathbb{F}$i ρ	F:$\mathbb{F}$i p	F:$\mathbb{T}$ ρ	F:$\mathbb{T}$ p
Confidence level	0.785	1e-50	-0.201	0.048	-0.799	1e-08
Time before hint	-0.196	0.054	0.209	0.040	0.124	0.225
Time after hint	-0.483	5e-07	0.201	0.049	0.456	2e-06
Hint proportion	-0.552	4e-09	0.286	0.004	0.493	2e-07
Full resolution	-0.149	0.146	0.165	0.107	0.091	0.374
Edited area	0.060	0.563	-0.151	0.137	0.004	0.965
Luminance level	-0.271	0.790	-0.598	0.558	0.095	0.351
Variability	0.019	0.850	-0.110	0.279	0.104	0.306

that the more challenging a forgery is, subjects will spend more time inspecting it and asking for hints, but for the truly hard ones nothing actually helps.

The results of our experiment suggest that there is *no correlation* between the relative size of the edited area and performance, contrary to our expectations. It is intuitive to think that the larger area the forgery covers on the image, the more likely for it to be spotted. The closest correlation found was to the class F:$\mathbb{F}$i with $\rho = -0.151$ and $p = 0.137 > 0.05$, which could suggest that smaller edited areas might scape subjects' attention. However, the null hypothesis could not be rejected in this case, either.

The F images consist of 20 *Erasing images*, 35 *Copy-Paste images* and 42 *Splicing images*. The average performance for each of those forgery types was 0.385, 0.469, and 0.594, respectively. These numbers *suggest* that Erasing images are harder to identify than Copy-Paste, which, in turn, are harder than splicing. While the number of images in each class is relatively small and unbalanced to support verification through a hypothesis test, these results make intuitive sense. Erasing forgeries have something removed from the image, so there is no outstanding element for the subject to look for. Copy-Paste forgeries duplicate a region on the same image, so there is an outstanding element, but its features (color, brightness, contrast) visually match the rest of the image. Splicing, on the other hand, inserts a completely foreign object in the image.

2.2.4 Anecdotal Observations

We collected and analyzed feedback from subjects to complement the results. This was obtained through a feedback form on the experiment's website, e-mail, con-

tact through instant messaging, and through face-to-face meetings. Less than 10% of the participants provided feedback, but the collected information provides insights on how subjects interacted with the study.

All replying subjects, including those with high performance, reported that they found the study hard in general. Subjects questioned in face-to-face meetings were shown some images from the test and inquired about their answers and justifications. From their responses, it appears that there is no consensus among subjects on what seems to be an *obvious* forgery. A similar lack of consensus was observed between the forensics experts that evaluated the level of difficulty of each image. Even the most explicit forgeries have eluded some subjects.

When missing a T image, by saying it was F, or missing an F image by providing wrong evidence, the most common justifications provided by the subjects were:

(i) The subject was fooled by a photographic artifact, such as lens flare, residue on the lens, or even by the photo's exposure;

(ii) After asking for a hint and spending a large amount of time analyzing the image, subjects felt compelled to guess that something had been manipulated;

(iii) Something in the context of the scene seemed wrong.

In turn, when missing an F image by saying it was T, the most common explanations were:

(i) The subject did not pay much attention to that particular part of the image, even after asking for a hint;

(ii) The image was too cluttered or there was too much in the image to be analyzed;

(iii) The subject did look at the manipulated region and found it suspicious, but plausible, or did not think it was a manipulation at all.

An important observation is that subjects rely strongly on context in order to make decisions. Several images depict people in social situations, and some individuals appear in multiple images. More than one subject reported being suspicious of particular characters after they thought that these characters have been used in forgeries. Some subjects went as far as to imply social relations, for instance assuming that two people depicted in the images were a couple and suspecting when one of them appeared with someone else. Images of cars and traffic also prompted contextual analysis by subjects. The most

notable occurrences were the cases in which a car was spliced driving in the wrong lane, and the one in which a Mercedes Benz logo was spliced over a Volkswagen vehicle.

2.3 Validation

Our paper is about how good humans are at perceiving image edits. Both humans and images are very diverse, and it is hard to find enough representative samples to make generalizations. Thus, a significant effort was made to find appropriate images and to validate the collected data. We divided the different types of validations in two categories: image dataset, and collected data. The first relates to the validation of actual images chosen for the test (Section 2.3.1), and the second to the pool of received answers (Section 2.3.2).

2.3.1 Dataset Validation

Our dataset consists of 80 T images and 97 F images. At first glance, these numbers might seem small, but the actual images were carefully selected to have good representativity. After performing the tests with the subjects, we further analyzed and validated our dataset in terms of image content and difficulty to better support our results.

2.3.1.1 Dataset Construction

Our first concern when selecting sources for our dataset was being able to compare the results of subjects with automatic forgery detection techniques. For this reason, we selected three datasets used by the image forensics community: *the forensics challenge database* (IEEE, 2013), the *splicing database* provided by Carvalho et al. (CARVALHO et al., 2013), and the *copy-and-paste database* by Cozzolino et al. (COZZOLINO; POGGI; VERDOLIVA, 2014).

In User Study Section of our paper we described a list of criteria used for chosing candidate images for our dataset. They were: *image type*, *image context*, *expected image difficulty*, and *edited area*. These criteria were judged subjectively at the time of the dataset construction. The manual process was divided into an initial selection and an iterative filtering.

Initial selection. In this step the first iteration pool was constructed. Every image

from the three databases was deemed either suitable or non-suitable for the test. The area and visual aspect of the edition were the most important factors in defining non-suitable images. Due to the nature of our test, only a single element (even if complex) of the image must be edited. Images that featured several spatially-disconnected changes were ignored. After this step the database contained around 500 images.

Iterative filtering: The images selected in the previous step were classified in 4 goups: true images, images containing splicing, images containing erasing, and images containing and copy-paste. The amount of images in each group was constantly accounted for balance. Over several iterations the groups would be inspected and a few images removed based on the mentioned criteria. If one of the groups had too many images removed, it would be out of the iterations until more images were removed from the others. The idea was to remove around 10 images from each group per iteration. We ended up with the 177 images used in the study.

2.3.1.2 Content Validation

To validate the distribution of content in our image database, we devised a series of tags describing elements of the scene. After that, all the images were manually evaluated and associated with one of more of the following tags:

1. *Indoor:* the image depicts an indoor scene;
2. *Outdoor:* the image depicts an outdoor scene;
3. *Natural Light:* scene lit by natural light;
4. *Artificial Light:* scene lit by artificial light;
5. *People:* there are people in the scene;
6. *Buildings:* there are buildings in the scene or the architecture of the building is a central element of the image;
7. *Nature:* there is vegetation or natural elements present in the scene;
8. *Vehicles:* vehicles are present in the image;
9. *House Objects:* objects that could be said to be household are present in the image;
10. *Work Objects:* objects that could be said to be related to work are present in the image;

11. *Street Objects:* street objects such as benches, lamp posts, fences, are present in the image;

12. *Landscape:* the image can be said to depict a scenic landscape, such as a mountain or beach;

13. *Day:* the image depicts a day scene;

14. *Night:* the image depicts a night scene;

15. *Signs:* there are signs, ads, billboards, or some sort of text or graphical element present in the image;

16. *Water:* a large body of water such as seas, lakes, or rivers can be seen in the image;

17. *Sky:* the sky can be seen in the image;

18. *Animals:* animals can be seen in the image.

Some of these tags are complementary, such as *day* and *night*. This means that an image can either depict a day scene, a night scene, or its undetermined, in which case none of the tags apply. Some tag combinations, while not common, are possible, such as having both natural and artificial light (an indoor scene with an open window and lights on).

Each one of the 177 images of the database was manually tagged for this validation phase, whose distribution can be found in Table 2.7 and Figure 2.7.

Figure 2.7: Visualization of the different combinations of tags in our image dataset. Each bar represents a different image. The stacked colors indicate different tags.

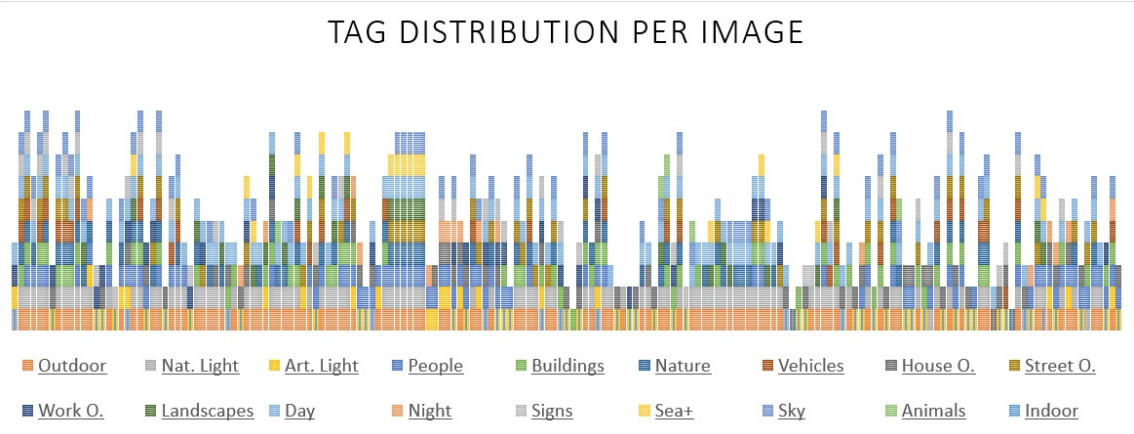


Table 2.7: Tag distribution among the images in our dataset.

Tag	**Images**	**Proportion**
Indoor	58	0.33
Outdoor	114	0.64
Natural Light	114	0.65
Artificial Light	62	0.35
People	81	0.45
Buildings	63	0.35
Nature	64	0.36
Vehicles	36	0.2
House Objects	53	0.3
Street Objects	55	0.31
Work Objects	28	0.16
Landscapes	27	0.15
Day	112	0.6
Night	18	0.1
Signs	41	0.23
Water	20	0.11
Sky	62	0.4
Animais	5	0.02

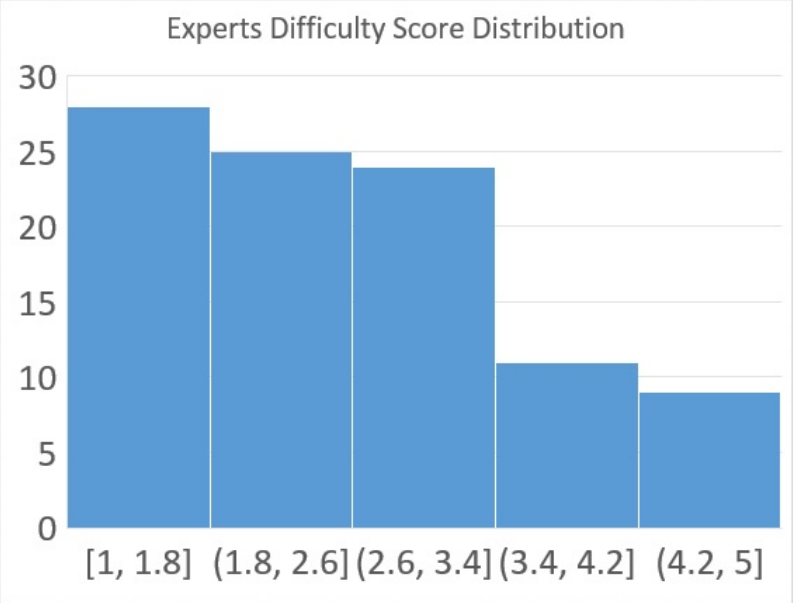
2.3.1.3 Accessing the Level of Difficulty of the Edits

To assess the difficulty of the forgeries in our database and provide further insight on the results, we performed a validation process with international forensic experts from FORLAB (FORLAB. . . ,) were asked to analyze each of the F images in our database and agree on a score for its difficulty. The score ranged from 1 to 5, with 1 being an obvious, easy to notice forgery, and 5 being a very hard to notice, well-done forgery.

The experts analyzed each image and tried to find the forgery in each of the images. After that, they observed the edited area and the users' evidence clicks, and discussed between them until agreeing on a score. Several images were already known to them.

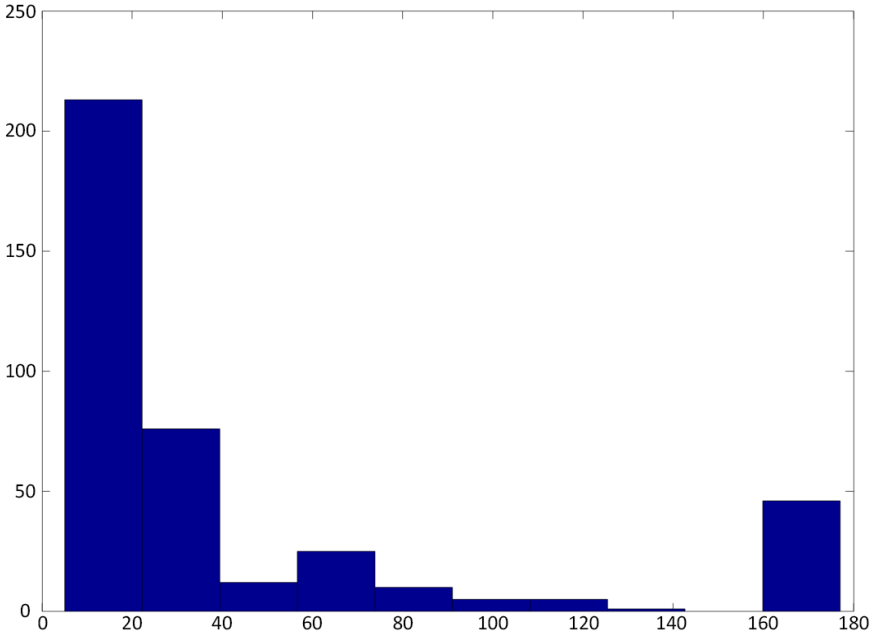
The nature of the analysis is deeply subjective, but there are no standard metrics for this type of evaluation. The experts considered how hard it was for the forgery to be spotted. The correlation between the performance of users and the scores was -0.46, with a p-value of $1.6e^{-7}$. This means that the experts' evaluation of difficulty matched the test results. According to the scores, the great majority of image edits were classified by the experts as easy to spot or at most having medium difficulty (Figure 2.8). This indicates that *even though the images in our dataset were not particularly challenging, the subjects performance was not good*.

Figure 2.8: Histogram of the scores assigned by two forensics experts for each image of our database according to the level of difficulty to detect the edits. A value of 1 means easy to detect, while 5 means very hard.



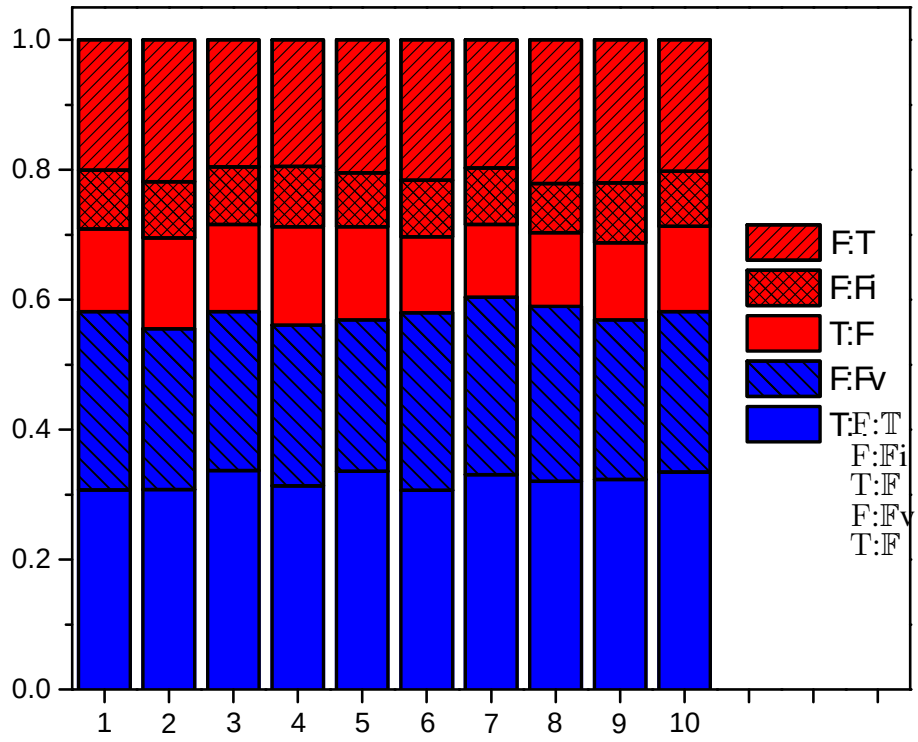
2.3.2 Data Validation

Figure 2.9: Amount of answers each subject provided. This X axis of the histogram represents an amount of answers, and the Y axis how many users provided that amount.



Due to the extensive size of our study, it is hard to guarantee that subjects analyze all images in the dataset. Only 24 subjects provided answers to all images, and there is a large variance in the amount of answers, and image types analyzed per subject (Figure 2.9). For this reason, it is important to determine if this causes some bias, and what is

Figure 2.10: Distribution of answers classes per subject level. This graph uses the current level of the subject at the time he provided a particular answer.



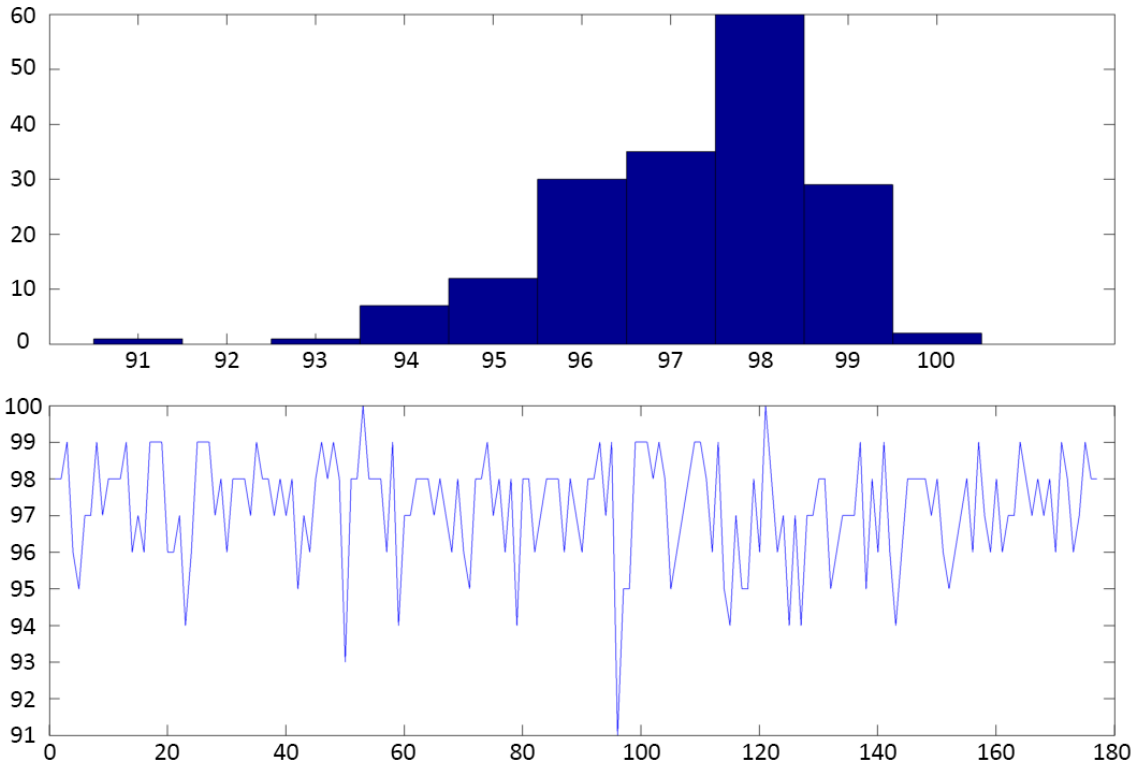
its extension, in the collected data.

2.3.2.1 Training Bias

The first thing to account for is the training bias. In essence, as they advance in the test, do the subjects becoming more experienced and does this change their performance? If so, this would require treating the subjects' answers differently as the test progresses. This is tracked by the subject level feature. Figure 2.10 shows the distribution of answer classes for all different subject levels. The bottom two classes represent the $\mathrm{T}{:}\mathbb{T}$ and $\mathrm{F}{:}\mathbb{F}\mathrm{v}$ classes, respectively, and their sum corresponds to the overall performance of that subject level. The subject level, in this case, is logged at the moment of the answer. The data in Figure 2.10 show that there is no significant change in the performance of subjects as they progress on the test, and the average performance is slightly below 60% for all levels. This is good for our methodology, because it allows us to treat all answers equally regardless of the subject's progress on the test.

The most important feature to reduce training bias was to only provide aggregated feedback. Subjects were only told of their performance after finishing a level, which usually consisted of analyzing approximately 15 images. There is no way to find out which images specifically the subjects answered right or wrong in the test, so it is not

Figure 2.11: Amount of answers received for each image. On top, the histogram of answers for all images, and on bottom the graph of answers received for each image. Notice the minimum amount of answers received by an image was 91.



possible for one to learn from his/her mistakes.

The algorithm for randomizing the image order was also responsible for reducing the influence of subject level. Every time a subject opened the test page, or answered an image and requested a new one, the 20 least answered images were determined. One of these 20 least answered images was then selected, at random. Since several subjects were online providing answers at the same time, the least answered images constantly changed. This was also important to guarantee all images had a similar number of answers, which is ideal for comparing data. The majority of the 177 images received between 96 and 99 subject answers, as can be seen in Figure 2.11.

2.3.2.2 Random Simulation

The requirement to provide evidence in the case of an $\mathbb{F}$ answer was specifically designed to avoid guessing. To test if this was achieved, we performed a random simulation and compared it with the obtained data. Each simulation consisted 100 answers per image, to be close to the reality of the test (Figure 2.11). A random answer consisted of a random $\mathbb{T}$ or $\mathbb{F}$ guess, and in case of an $\mathbb{F}$ guess, a random position on the image was

also selected. The answers were then evaluated using the same classification criteria as the subjects'.

The obtained distributions after 1,000 different simulations is summarized on Figure 2.4. The simulations were equivalent to the implicit probability calculation for each class. For example, the implicit probability for the $\mathrm{T}{:}\mathbb{T}$ class is $p_{\mathrm{T}:\mathbb{T}} = p_{\{\mathrm{T}:x\}}p_{\{x:\mathbb{T}\}}$, where $p_{\{\mathrm{T}:x\}}$ is the probability of the image being T, and $p_{\{x:\mathbb{T}\}}$ is the probability of the random answer being $\mathbb{T}$. In our data, $p_{\mathrm{T}:x}$ is approximately 0.45 (the percentage of all true images is 45%), and $p_{x:\mathbb{T}}$ is exactly 0.5 (either $\mathbb{T}$ or $\mathbb{F}$), meaning $p_{\mathrm{T}:\mathbb{T}} = 0.225$, matching closely to our simulation results.

The probability of simulated guesses for $\mathrm{F}{:}\mathbb{F}\mathrm{v}$ answers, described as $p_{\{\mathrm{F}:\mathbb{F}v\}} = p_{\{\mathrm{F}:x\}}p_{\{x:\mathbb{F}\}}A_e$, where A_e is the percentage of the image area accepted as evidence, is extremely low. This makes it very hard for a random simulation to correctly guess edited images in our test. In this case, the simulation should guess the image and the answer are F and $\mathbb{F}$, respectively, but would also need to correctly guess an evidence position, which must fall inside the actual evidence area. In our tests this is very rare, with $p_{\{\mathrm{F}:\mathbb{F}v\}} < 0.03$.

2.3.2.3 Balanced Resampling

Besides our answer pool being unbalanced in the amount of answers provided by each participant, the random nature of the image-selection process also accounted for an imbalance in the number of answers per image type. As discussed in the *Image Statistics* Section of our paper, if the different types of images present different difficulties to subjects', this could bias the final result towards the most common type of forgery. To check the occurrence of this type of bias, we performed a resampling experiment, selecting balanced subsets of data from the total answer pool and comparing them with the complete set. For each subject, we counted the types of forgeries they answered (splicing, copy-paste, and erasing). The minimum amount answered is used a reference to create a balanced subset of answers from this subject. Thus, if the person answered 4 erasing images, 10 copy-paste, 15 splicing, and 40 T images, the minimum value is 4. We then select the 4 erasing images, and randomly pick 4 copy-paste and 4 splicing images answered by the participant, for a total of 12 F images, to for this participant's balanced subset of answers. This is then also balanced with 12 T images. This will cause 6 copy-paste, 9 splicing, and 28 T images to be disregarded in this validation experiment. We can resample this subject's answers alone for a total of $\binom{10}{4}\binom{15}{4}\binom{40}{12} = 1.6 \times 10^{15}$ different balanced subsets, meaning there is a massive number of possible resamplings.

Each complete pool of balanced answers following this criteria will have 10,051 answers. Note that if a subject answered 80 T images, 30 splicing images, 20 copy-paste, but not erasing, his minimum will be 0, and all his answers will be discarded. This is statistically unlikely, however.

To compare with both the random simulation and our complete answer pool, we performed 1,000 different resamplings varying images from the answer pools of all subjects. The average class distribution among all resamplings can be seen in Figure 2.4. There is almost no different on the obtained results, except for a little increase in T images answered. This can be explained as the original proportion of T to F images is 0.45 to 0.55 in our database, while in the balanced case it is 0.5 to 0.5. *This experiment shows there is no significant bias inherent on the imbalance of image types in our data.*

Table 2.8: Correlation and respective p-value between different image features for all images. Here, blue denotes positive correlation with acceptable p-value ($p < 0.05$), and red denotes negative correlation with acceptable p-value. Black values do not satisfy the threshold and the null hypothesis cannot be rejected.

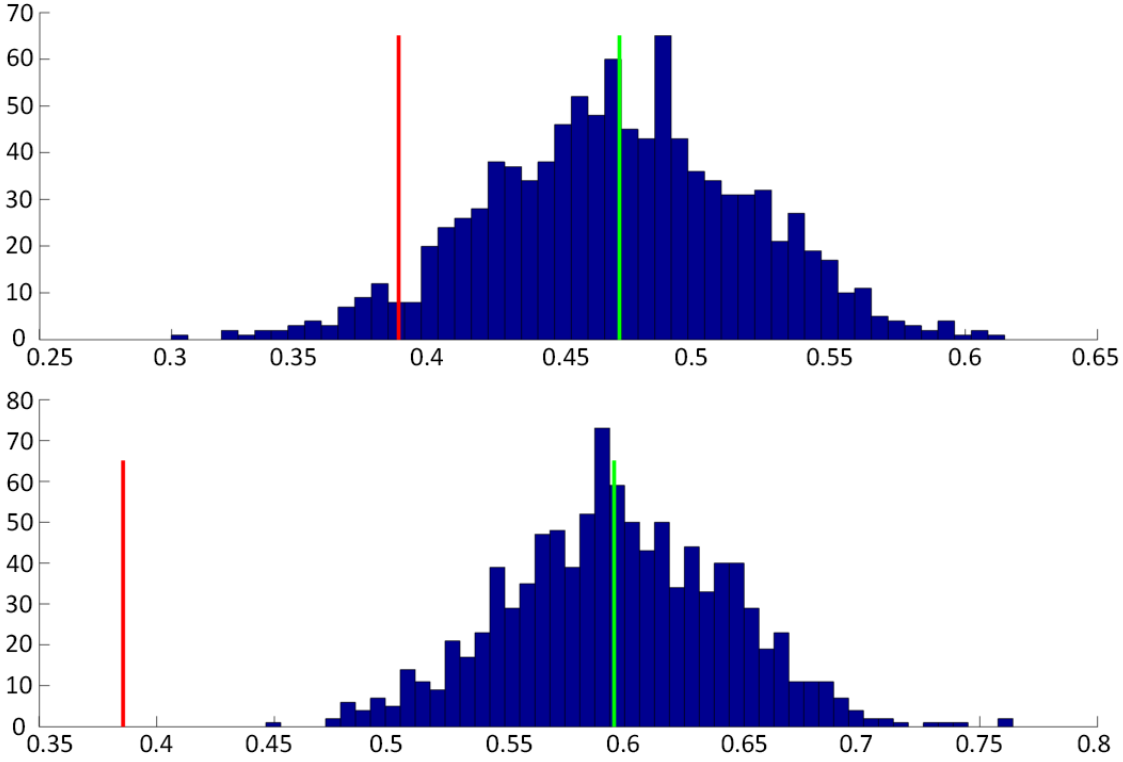
	C ρ	C p	T ρ	T p	TH ρ	TH p	H ρ	H p	FR ρ	FR p	L ρ	L p	V ρ	V p
P	0.200	0.007	-0.098	0.192	-0.119	0.114	-0.094	0.212	-0.026	0.727	-0.045	0.556	0.051	0.500
C	-	-	-0.245	0.001	-0.579	3e-17	-0.684	9e-26	-0.221	0.003	-0.001	0.992	-0.149	0.048
T	-	-	-	-	0.384	1e-07	0.337	4e-06	0.037	0.627	-0.030	0.691	0.118	0.117
TH	-	-	-	-	-	-	0.764	3e-35	0.252	0.001	0.054	0.479	0.374	2e-07
HP	-	-	-	-	-	-	-	-	0.205	0.006	0.073	0.335	0.169	0.025
FR	-	-	-	-	-	-	-	-	-	-	-0.094	0.212	0.177	0.019
L	-	-	-	-	-	-	-	-	-	-	-	-	0.154	0.041

Our dataset presents a different amount of splicing, copy-paste, and erasing forgeries (42, 35, and 22 respectively). This is a small and unbalanced amount of samples to assert any hypothesis about differences in these three classes. Nevertheless, we performed a resampling analysis to analyze the change in subjects' performance between them (Figure 2.12). We used erasing as the basis of comparison as it was the smaller set, and resampled 2,000 different subsets of 22 images with Copy-Paste (top), or Splicing (bottom) forgeries. The obtained distrubition in average performance is then compared with that of erasing forgeries. For the gathered data, all sets of 22 spliced images had a better average perfomance than for erasing forgeries, and the great majority of Copy-Paste ones. This is commented in the results Section of our paper.

2.4 Related Work

To the best of our knowledge, ours is the most comprehensive study to evaluate people's ability to identify false images. It stands out mainly due to its extension, and the

Figure 2.12: Results of the performance distribution on the resampling process for Copy-Paste (top) and Splicing (bottom) compared to Erasing images. The green line represents the average performance for all images of that type (35 for Copy-Paste and 42 for Splicing), while the red line is the average performance for Erasing forgery.



introduction of evidence analysis (Figure 2.2). There have been, however, related studies with different scopes and scales. In the field of *perception*, Ostrovsky et al. (OSTROVSKY; CAVANAGH; SINHA, 2005) explored how different lighting configurations influence user perception. The study explored different forms of visual stimuli: 3D computer-generated scenes, photographs, and pictures were shown to the subjects. The authors measured the response time and accuracy in detecting lighting irregularities in different configurations. The study concluded that it is easy to assess outliers in a small set of objects, but the task becomes very hard in complex scenes with different objects and light interactions. Unlike our study, this one systematically evaluated the perception of a specific aspect (*i.e.*, lighting) using a small number of subjects (17) in a strictly controlled environment.

Farid et al. (FARID; BRAVO, 2010) studied users' performance at detecting irregularities in geometry, shades, and reflections, and discussed its forensics implications. For this purposes, the tests involved twenty subjects observing different pictures and trying to identify tampering. Their results show that humans are inept at perceiving inconsistencies in shadows, reflections, and planar perspective distortions. Furthermore, forensics

solutions were presented that could be used to help users identify these types of forgeries. This study differs from ours by focusing on the perception of specific geometric features, and by using a small group of subjects.

Another work by Farid et al. (FARID; BRAVO, 2012) assessed the subjects' ability to distinguish photographs of human faces from computer generated (CG) faces. The results show that while humans can reliably distinguish photographs from CG models under various circumstances, modern CG techniques and good 3D modeling pose very hard challenges. A recent study by Holmes et. al. (HOLMES; BANKS; FARID, 2016) used more advanced 3D renderings of portraits. The subjects performed poorly at identifying CG portraits of people, but the authors showed that with a small amount of training much better results can be obtained.

Ghadiyaram and Bovik (GHADIYARAM; BOVIK, 2016) perfomed a massive crowdsourced online study to evaluate picture quality. Their objective is to measure the effect of common distortions on images, such as capturing a photo with a mobile device, compression, noise, etc. on human perception. They constructed a dataset of 1,162 images with varying degrees of distortion, then gathered user opinions on image quality. Using a crowdsourcing platform, they were able to gather over 350,000 opinions, with over 8,100 different observers. They validated their data after for answer consistency, and tested automatic algorithms for image quality assessment. This work has similar approach to ours, and is faced with similar challenges of a large-scale subjective study, which the authors discuss in the paper. Since their objective is different from ours, a direct comparison of results cannot be made with our work. For instance, they use no ground truths that can be used to estimate accuracy.

Carvalho et al. (CARVALHO et al., 2013) proposed a technique to detect forgery based on color classification of scene illuminants. To validate their approach, the authors created an image database with true and spliced images, and performed a variety of tests, some of which involved human subjects. Similar to our study, theirs did not focus on any particular aspect of human perception. Their study included approximately 2,000 individual answers (ours used over 17,000 answers) over 200 images. The reported accuracy of the tested subjects was of 64.7%, identifying only 38.3% the false images. The authors used a binary classification (true or false images), implying that the accuracy could have been influenced by subject guessing the right answer based on incorrect premises.

A more recent study published after our paper's submission performed a very similar experiment to ours. In "Can people identify original and manipulated photos of real-

world scenes?" (NIGHTINGALE; WADE; WATSON, 2017), the authors created a set of manipulated images from 10 base images, and tested on around 700 subjects. Each person viewed ten images, five of which were originals, and the other five were edited. The authors focus on the visual and cognitive aspects of noticing each forgery, due to its effect on the image. In their experiment, classifying the image as "real" or "manipulated" was a different task than locating the image, as we do with our evidence. The results for both their tasks had remarkably similar accuracy to our experiments, around 60% for classifying (our classification results can be estimated from Table. 2.2), and 56% for location.

All of these works support our findings that humans are not generally good at identifying forgery in digital images.

2.5 Summary

We have performed a study to evaluate the ability of an average individual to spot edited images. The subjects in our experiment were mainly young adults, with college-level education, and at least some experience with digital images (Table 2.4). This is one of the most well-suited demographics for achieving good performance for this particular task. According to an OECD[3] study (OECD, 2016) that measured the skills of adults in more than 24 countries, we can argue that our subjects represent the upper bounds of the general population in terms of education and technological skills. Thus, it seems reasonable to argue that the results observed in our experiment actually correspond to a higher performance rate than the average population.

Our findings, supported by statistical evidence, suggest that humans are likely to be fooled by digital images at least half of the time. Participants not only missed identifying forgeries in images, they also often questioned the authenticity of pristine pictures. Our study also demonstrated that the nature of an image and its features may affect ones ability to detect forgeries. As such, further work is required to better understand the relevant aspects involved in such observed behavior.

[3] Organisation for Economic Cooperation and Development (OECD)

2.5.1 Main Findings

The core results of this study showed that humans were capable of providing correct answers to whether an image has been edited or not only 58% of the time, and were capable of identifying actual forgeries in only 46.5% of the cases. Experience and young age influenced positively the results. The subject's behavior (time, hints, confidence level, etc.) during the study had the biggest impact on the success rate. *Since behavior can be taught, this suggests that there might be strategies and good practices to aid subjects in spotting modified images.* This is corroborated by Holmes findings (HOLMES; BANKS; FARID, 2016), and should be verified in further studies.

These findings are of special importance to the forensics community. They show that we are not able to tell if the content we are being exposed to is truthful. Be it a friend editing their social media pictures or a news agency enhancing their photographic reports, people have a low chance of spotting them on their own. In fact, recent research has shown that even if someone believes an image has been edited and searches the web to gather evidence, it is not easy to prove it (CONOTTER et al., 2014). This outlines the importance of having tools to help people analyze and authenticate images.

2.5.2 Limitations and Further Investigations

The subject discussed in this Chapter deals with both people and images, two things that have a complex nature and great variability. No amount of participants or test images could be enough to provide definitive answers on this subject. Nevertheless, we performed the largest study of this kind so far, and validated our test dataset, methodology, and collected data.

Given the nature of the study and the amount of uncovered data, several things can be further investigated. For example, there are 8,160 points of evidence provided over the 177 images in the form of subject clicks. This data could provide insights on what kinds of objects or image elements the subjects are more prone to suspect.

It is possible to test the influence of different image composition techniques on the quality of the forgery. A *splicing* forgery, for instance, can be done by simply cutting and pasting a region from an image into another, or by using sophisticated tools. Alpha Matting (GASTAL; OLIVEIRA, 2010), and gradient domain composition techniques (DARABI et al., 2012) are able to blend two images, creating visually imper-

ceptible compositions. They are not perfect, however, and differences in perspective or illumination may reveal them. A study using our methodology focused on different types of *splicing* forgery could help us better identify dangerous composition techniques.

Another important front of research is relating our results with works on public perception of digital images. Conotter et. al. (CONOTTER et al., 2014), for instance, performed a large-scale study on subjects perception of altered images. They showed side-by-side pairs of true and altered images to a large group of individuals and asked them to rate the level of alteration and how they felt the changes affected their perception of the images. It could be interesting to apply our testing methodology using their dataset to new subjects and look for correlations between the reported level of image modification and difficulty to detect it.

On the same line, Ghadiyaram works on perception of image quality (GHADIYARAM; BOVIK, 2016) can be combined with our approach to better understand the effect of distortions. We showed that people can be tricked into suspecting there is something wrong with an image due to artifacts, so it is intuitive that image quality affects forgery localization. User evidence maps and answers can be combined with models for quality prediction (LIU et al., 2016), both improve prediction and to better understand the effect of distortions on perception.

The main goal of this work was to develop a proper basis to ground forensics research, by collecting factual evidence on the perception of image forgery by users. What we have uncovered shows that humans are, indeed, easily fooled.

3 DIGITAL IMAGE FORENSICS VS. IMAGE COMPOSITION

Our study on human perception (Chapter 2) highlights the importance of tools to assist people to perform image analysis. Following these results, Chapters 3 and 4 explore the practical potential of such tools. This Chapter reproduces the content of the article *Image Forgery Detection Confronts Image Composition* (SCHETINGER et al., 2017a), adapted for the thesis' context.

In this work, we argue that DIF is technically advanced to detect most types of forgeries. To demonstrate that, we analyze the "arms race" between the fields of image forensics and image composition. Here, *image composition* is used as an umbrella term for all techniques from areas such as computer graphics, computational photography, image processing, and computer vision that could be used to modify an image. More specifically, we discuss research that has the potential to either be used to perform or hide forges in digital images.

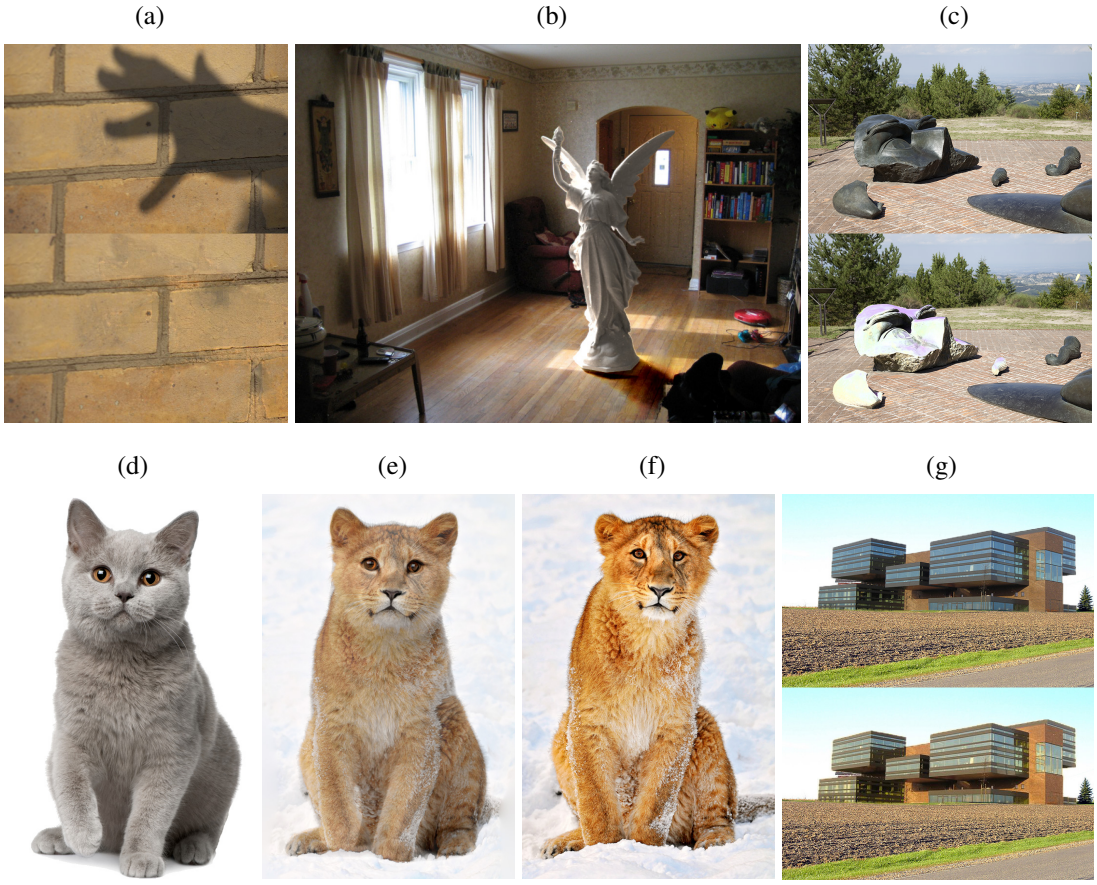
From a forensic point of view, image manipulation is usually classified as either *splicing*, *copy-pasting* (also called *cloning*), *erasing*, or *retouching*. A splicing operation consists of transferring an object from an image into another, but this could be done simply by cutting and pasting, or using an advanced technique that matches the gradient of the target image (PÉREZ; GANGNET; BLAKE, 2003). Copy-pasting is similar in essence, but the transferred object comes from the same image. Retouching has a vague definition that could fit a wide range of actions, such as blurring regions of the image, recoloring, and applying filters (BHARATI et al., 2016).

While many modern image composition techniques could be used to make sophisticated forgeries, almost none of them have been scrutinized by forensic literature. There is, however, a large body of forensic tools that could be used for this task. This work surveys both the forensics arsenal and image composition techniques to identify the best strategies to analyze these novel forgeries. Our main objective is to provide a starting point for researchers in either field that want to venture into their counterpart, focusing on the forensics point of view.

The main **contributions** of this chapter include:

- The introduction of a Forgery Detection scale for forensics assessments (Section 3.1), which classifies the output of forensics techniques according to the type of information uncovered. It is also used to extend existing classifications of forensics techniques (Section 3.2);

Figure 3.1: Different examples of composition techniques used to alter images. (a) Removing soft shadows (GRYKA; TERRY; BROSTOW, 2015): the hand shadow from the top image has been removed in the bottom image. (b) Inserting synthetic objects (KARSCH et al., 2014): the marble angel in the picture is not real, it was rendered along with its complex light interactions. (c) Performing edge-aware filtering (GASTAL; OLIVEIRA, 2015): the bottom image was filtered to perform a localized color editing in some of the stone statues. (d)-(f) Morphing two different objects together to create a blend (LIAO et al., 2014). The cat in (e) is a composite of the cat in (d) and the lion in (f). (g) Transferring an object from one image to another, adjusting its illumination according to the target scene (XUE et al., 2012a): the building was spliced on the field in the top image, and in the bottom it had its lighting adjusted to match the composition.



- A review of the state-of-the-art in Image Composition from a forensics point of view, organizing techniques by their forgery potential, and discussing their strengths and weaknesses against detection (Section 3.3);

- A qualitative analysis of several composition techniques to an uncontrolled image set, showing promising interactions with traces for exploration (Section 3.4.1);

- A quantitative analysis on a fully controlled image set, providing an in-depth analysis of the behavior of double JPEG compression on different composition techniques (Section 3.4.2).

3.1 The Forgery Detection Scale

Forgery detection is concerned with determining if, where, and how, a target image has been modified. This is achieved by analyzing different traces. Traces refers to any residues or marks left on an image. The image acquisition process introduces traces due to the scene lighting, the camera lens, the digital sensor, and so on. Image compression or editing leave traces or alter pre-existing ones. The image itself can be said to be essentially a big collection of traces. Forgery detection then works by either looking for patterns in traces where there should not be any, or by looking for the lack of patterns where there should be some.

Consider, for instance, an object spliced from an image into another, and re-sized to fit the target picture. Resizing an area to be either bigger or smaller will involve re-sampling the original pixels. This creates a correlation between neighboring pixels, introducing a new pattern that was not originally there. If the target image was JPEG compressed, the local block structure would be disrupted by pasting something over it. This breaks a pattern that should be there. Identifying and measuring traces is a challenging task, as successive operations can modify or destroy them. The task of a forensics analyst is to investigate the image for traces, with the help of forensics tools and techniques.

The most simplistic view would pose that there are only two outcomes for forgery detection: the image has been altered, or no evidence of alteration is found. However, this classification might not be sufficient. Simply compressing an image might be considered an alteration, even though it is a commonplace operation, making this classification useless. Different forensic techniques work on different assumptions of what traces could be present in the image and what it can be inferred from them, e.g., the location or the nature of the forgery. There is no standard in the literature, however, for classifying and comparing techniques based on their outcomes.

Here, we propose a new general classification scale called the FD (short for Forgery Detection scale). **This scale evaluates the output of forensics techniques, and the information uncovered from it**. This scale is based in the concept of an image being native or not: a native image is an image that was captured by a device and then outputted to the user “as-is". Conceptually, this is easy to define, but technically there might be some complications: different devices process the image differently.

Consider a technique that estimates the level of compression the image has undergone. If the image is in the PNG (Portable Network Graphics) format, any level of

compression estimated can be used to put its nativity in question, because it is improbable for a native .png image to have compression. If we are analyzing an image from a known device, we might know the processing pipeline and the amount of compression to be expected. In that case, finding no traces of compression or finding a different level of compression would mean the image is not native. These problems will be discussed in detail further in the paper.

The FD scale ranks forensic techniques based on the *type of evidence* they can possibly provide about an image's history. It does not rank the image itself. The first possible outcome is the negative case, when it is not possible to discover information supporting that the image has undergone any modification with respect to what is known about it. This could happen because the image is actually native, or because the analyzed traces do not show forgeries. There is no practical difference between these two cases: it is not possible to say that an image is truly native, only that there is no evidence supporting that it has been modified. This outcome falls outside our scale in practice, but can be called FD0 for convenience. The following are the different levels of our Forensics Detection scale:

FD0 No evidence can be found that the image is not native.

FD1 The image has undergone some form of alteration from its native state, but the nature and location of it is unknown;

FD2 The image has undergone some form of alteration from its native state, and its location can be narrowed down, but no information about its nature was obtained;

FD3 The image has undergone some form of alteration from its native state, there is information both about the location and its nature;

FD4 All the conclusions of the previous item, and there is information to link a particular processing tool or technique to the alteration.

These should be referred as FD (short for forgery detection) scale, with values FD0-4. It could be argued that the ultimate form of forgery detection would go beyond identifying the used technique, such as identifying the forger. Although this could be a desirable point, it is currently not common in digital image forensics, and falls beyond the scope of this work. Research in image phylogeny (DIAS; ROCHA; GOLDENSTEIN, 2012) proposes to estimate the history of the image's alterations or different versions over

time, and the FD scale applies. If different versions of an image can be accounted for the phylogeny process, at least one of them cannot be native (FD1), and by comparing the differences among them one can uncover where (FD2) and what (FD3) has been changed.

The FD scale is backwards inclusive for FD>0, meaning that if FD4 can be guaranteed so can FD3, FD2 and FD1. The following subsections provide in-depth explanation of the different levels of FD scale and further considerations.

3.1.1 FD0: Non-meaningful Evidence

Whenever any form of inspection is done without informative results, it can be said to be FD0. This stems from the fact that in forensics there is no evidence of non-manipulation. By accumulating evidence, one can show that an image has been altered, but no amount of inspections can guarantee it has not been altered. In this sense, FD0 does not depend on the nature of the traces analyzed (coding, compression, editing), or the sophistication of the technique. Whenever an inspection is done without providing additional information to the analysis, we define it as FD0. This can happen because the image is genuinely pristine, or due to a false positive response.

3.1.2 FD1: Nativity Information

The FD1 level differs from the negative case FD0 because it is possible to determine that the image is not native. This is not so simple to assess, as most modern cameras have a processing pipeline comprised of several operations (demosaicing, white balance, etc), changing the image before the user has access to it. Furthermore, demosaicing is such a fundamental operation in modern cameras that it makes little sense talking about images without it. For the sake of generality, we propose that any form of pre-processing in the image up until a single in-camera compression can be accepted without breaching the image nativity. In cases where there is information about the image's history or pre-processing, this definition can be adapted. When considering images acquired from social networks or photography apps, their native states might contain many operations, such as filters, compression, resizing. A forensic technique achieves FD1 when it is able to find evidence of alteration after capture. Techniques that analyze an image's EXIF (exchangeable image file format) information are an example of FD1: they can detect an

inconsistency in the metadata proving an image is not native, but nothing can be said about location or nature of the alteration. This is similar to ENF (Electric Network Frequency) analysis (COOPER, 2008), where a known behavior or standard is used for comparison. In the case of ENF Analysis, the power line frequency is used as a watermark. For images, known processing pipelines for digital cameras or social networks can be used in a similar way.

3.1.3 FD2: Location Information

The FD2 level is obtained when the general location of the alteration in the image is known. It is possible that a region of an image has been erased by a series of copy-pasting operations, and then retouched with smoothing brushes. In this case the boundaries of the forgery might not be as clear. If a technique is able to obtain any form of specificity in the altered region, FD2 is achieved. This is the case when analyzing traces such as PRNU (Photo Response Non-Uniformity, Section 3.2.1), CFA (Color Filter Array, Section 3.2.1) or ELA (Error Level Analysis, Section 3.2.2), that are locally structured in the image. Many forensics techniques generate an image as output, or an output map, which represents some local measure. It can be a binary mask, a probability map of pixels having a certain property, or any spatial form of measure. Since any information gathered in this form is locally constrained, it will be at least FD2.

If evidence of any global alteration in the image is found, then the location of the forgery is the whole image. Similarly, operations that remove parts of the image such as seam carving and cropping can be detected but the actual altered area is not present in the analyzed image anymore. It is argued that the forgery location can be considered to be all image, reaching FD2.

3.1.4 FD3: Nature Information

When the output of a forensic technique can help understand what has happened to an image, it is considered FD3.

The nature of the forgery can be subjective, because it is not possible to predict all ways in which an image can be altered. The most commonly studied forms of forgery such as splicing, copy-pasting and erasing, are just a subset of possibilities. As was discussed

on the introduction, image composition techniques are able to alter the shape, texture and orientation of objects, and even merge them together. For simplicity, any meaningful information in addition to location of the processing that can be used to assist the forensics analyst can be considered FD3. For instance, identifying that a spliced object has been rotated and scaled awards an FD3 level on the scale. Even identifying that an object is just spliced is worth an FD3 on the scale because the image is not native (FD1), its location on the target image is evident (FD2), and the nature of the alteration is known (FD3).

3.1.5 FD4: Technique Information

The highest level in our scale, FD4, is achieved when the analyst finds evidences that can link the forgery to a particular technique or tool. A splicing can be done by simply cutting a region from an image an pasting over another, but there are also sophisticated ways to blend them, such as Alpha Matting or Seamless Cloning. A forensic technique that is able to, after obtaining FD3, provide further insight into the technique or tool used to perform the forgery achieves FD4. Achieving FD4 is a challenging task, as knowledge of digital images is not sufficient: particular techniques must be known and understood. The feasibility of this level depends on the type of manipulation, and it is possible that it is unachievable in many cases.

3.1.6 Accuracy and Confidence

The purpose of the FD scale is to understand what type of information can be uncovered in a certain context. Forensics techniques have different outputs, and certain operations on images also destroy information. **The scale does not evaluate the accuracy of techniques, their confidence or applicability.** It is a qualitative measure that needs contextualization, and it is not a metric that can be automatically determined in a simple way.

If a technique provides an output map of irregular pixels based on a general trace such as PRNU, it is going to be FD2. Two different techniques that produce the same type of maps based on PRNU, but one has better results are still both FD2. A forensic technique that outputs the same type of probability map per pixel, but is looking for traces left by a particular processing like local gaussian filtering would be an FD3 in our scale. Visually

both maps could be similar, but they are providing a very different type of information.

3.1.7 Black-Box Approaches

Some forensics approaches rely heavily on feature descriptors or machine learning to strictly classify images as having some type of forgery. For instance, evaluating an image as a whole trying to identify if it has been spliced, but without providing a location. We refer to these as “black-box" approaches, because there is a layer of abstraction encapsulating the inner workings, and information is obscured.

At first, the technique described in the example seems to fall in FD3 but not in FD2. Our argument is that such general techniques are actually testing for the presence of traces, that could be introduced or altered in several different ways: detecting splicing in this sense is only detecting non-nativity (FD1) as one is unsure how the technique responds to other types of forgery. Unless the authors exhaustively tested for other types of forgeries, it is not possible to assume that there is an intrinsic separability among forgery classes. For instance, a technique for detecting splicing might also respond positively to copy-paste or erasing. This might not be an issue if the technique provides information regarding its decision process. When using convolutional neural networks, for instance, it is possible to visualize the activated layers that triggered a positive response for an image, providing location information (FD2).

We argue that black-box approaches should be avoided. Firstly, there is a reproducibility issue regarding implementation and datasets. Even if the code is provided, having an obscure layer of abstraction limits its usefulness in the forensics context. For many real cases of analysis, such as investigation, courts of justice, journalism, etc. articulating the result of the analysis is crucial. Having a technique that is easily explainable or usable in an argument is more useful than a black-box that has achieved perfect results on its tests. Some techniques might not be black-box per choice, as experimentation can yield positive, publishable, results prior than complete understanding. In Section 3.5 we discuss research such as standard datasets, and decision fusion, that can help in avoiding “black-boxing".

3.2 The forensics arsenal

The current state-of-the-art in digital image forensics provides an arsenal of tools and techniques for forensics analysts. Here we investigate the most relevant approaches and their capabilities, both in terms of applicability (i.e. when we can use them) and assessment (i.e. the level that can be achieved in FD scale). In a general way, it can be noted that there is a trade off between the generality and the FD level that a technique is able to reach. This is intuitive, because the higher the level on the scale, the more specific the assessments are. FD1 can be simplified as a boolean statement (the image is either native or not). From FD2 onwards, there is a large set of possible answers (all different combinations of pixels in the image). To identify the nature of the forgery (FD3), a technique must be looking for more specific features or traces.

An image forensic tool is usually designed considering 3 steps:

1. Some **traces** in the image - possibly introduced by the forgery process - are identified; such traces can be scene-level information such as “the image lighting", or signal-level, such as “the color filter array pattern".

2. These traces are measured and quantified in some way, resulting in **features**, which are usually numeric in nature;

3. By analyzing through experimentation how the set of features behaves for native and forged images, a **decision** is taken about the image. This can be done using simple thresholds or sophisticated machine learning techniques.

Table 3.1: The steps of the tool by Carvalho et. al. (CARVALHO et al., 2013).

Layer	**Example**
Trace	Illuminant or light source of the image.
Feature	Estimated illuminant colors and light intensity on object edges.
Decision	SVM classification.

In Table 3.1 we show a practical example of previous steps for the technique developed by Carvalho et. al. (CARVALHO et al., 2013) to detect splicing. The used **trace** is the illuminant, or the light source. The key observation is that if an object is spliced and the original image had different light conditions, such as indoor or outdoor lighting, or

even incandescent vs. fluorescent lights, this trace can be used to identify it. The **features** used are the estimated illuminant colors and the light intensity on the edges, for the different analyzed regions of the image. The **decision** process uses a Support Vector Machine (SVM) to classify the image as either spliced (FD3) or inconclusive (FD0) based on the features.

Forensic tools classification is based on the traces they analyze. Piva (PIVA, 2013) distinguishes between traces left by three different steps of the image formation process: acquisition, coding and editing. Another intuitive classification has been proposed by Farid (FARID, 2009b) where the forensic techniques are grouped into five main categories: pixel-based, format-based, camera-based, physically-based and geometric-based. Farid's classification is more common in the literature and distinguishes better between traces, but Piva's can be closely related to the image history. We propose a classification based on Piva's approach (Fig. 3.2), but with greater specificity, similarly to Farid's. The FD scale will be used to describe which level of assessment can be expected when examining an image using a specific tool.

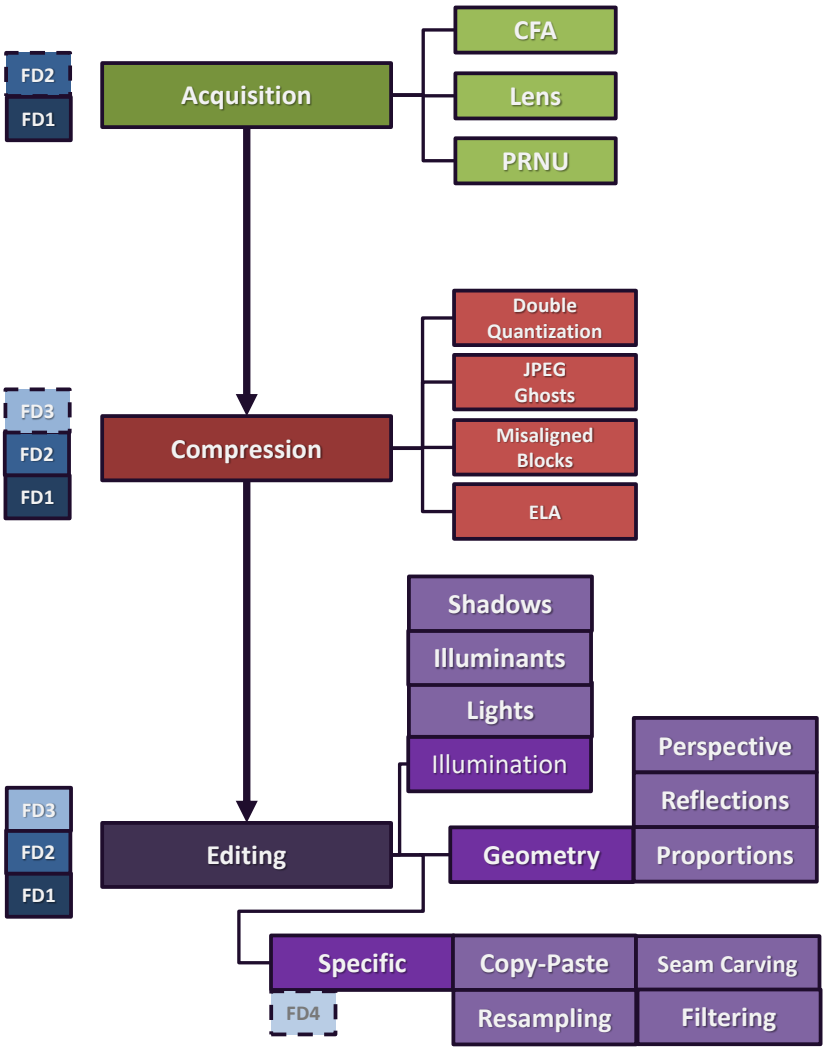
3.2.1 Acquisition Traces (AT)

Native images come to life with distinctive marks (artifacts, noise, inconsistencies) due to the acquisition process. Both hardware (e.g., lens, sensor) and software components (e.g., demosaicing algorithm, gamma correction) contribute to the image formation, introducing specific traces into the output (native) image. When a native image is processed some of these traces can be deteriorated or destroyed, exposing evidence of tampering. Forgery detection using acquisition traces generally falls in one of two categories:

1. *Global*: The analyzed trace is a global camera signature. Non-native images can be exposed when this signature does not match with the supposed source device. For instance, in (KEE; JOHNSON; FARID, 2011) non-native JPEG images are exposed by analyzing quantization tables, thumbnails and information embedded in EXIF metadata. In (LUKAS; FRIDRICH; GOLJAN, 2006) the reference pattern noise of the source device is taken as a unique identification fingerprint. The absence of the supposed pattern is used as evidence that the image is non-native.

2. *Local*: The analyzed trace has a local structure in the image. Its inconsistencies in some portion of the image can be exploited to localize the tampering. For instance,

Figure 3.2: Forensic techniques' classification. Each type of trace is organized under its correspondent phase in the forgery process. The techniques themselves were omitted for the sake of clarity, but would appear as leaf nodes under their analyzed traces. On the left, the relation to the FD scale is displayed. Only by analyzing specific editing traces it would be possible to achieve FD4.



Ferrara (FERRARA et al., 2012) uses demosaicking artifacts that form due to color interpolation. They can be analyzed at a local level to derive the tampering probability of each 2×2 image block. Chen (CHEN et al., 2008) reveals forgeries by detecting the absence of the PRNU on specific regions of the investigates image.

Let us note that some traces can be considered both at a global or local level (e.g., demosaicing artefacts and PRNU), allowing to identify non-native images (FD1) or to localize forgeries (FD2). The analysis of acquisition traces is usually limited to matching a known pattern, and they can be easily disrupted. For this reason, FD2 is the highest we can expect to achieve on the FD scale using acquisition traces. The analysis of

acquisition traces generally requires some additional information about the source device. In some cases this information depends on the source device model or manufacturer (e.g., color filter array pattern, quantization tables), and can be easily obtained to assess image nativity (HASS, 2017). In other cases these traces are unique camera fingerprints (e.g. PRNU) and can be obtained by having the source device available, or be estimated by using different images captured by the same device.

3.2.2 Coding Traces

Lossy compression might happen in many occasions during the life of a digital image: native images of non professional cameras and smartphones usually come to life in JPEG format; when uploading a photo on a social network lossy compression is usually applied to the image; when a JPEG image is altered and saved again in JPEG, double lossy compression occurs. This is called double JPEG compression, or D-JPEG for short. For this reason, most of the literature has focused on studying the traces left by single and multiple chains of JPEG-compression. This is a very prolific field of study in forensics, with a wide variety of techniques. Fan (FAN; QUEIROZ, 2003) and Luo (LUO; HUANG; QIU, 2010) provide efficient methods to determine whether an image has been previously JPEG compressed, and, if so, are able to estimate some of the compression parameters. Further advances have been also provided by Li et al. (LI et al., 2015) to identify high-quality compressed images based on the analysis of noise in multiple-cycle JPEG compression. On Bianchi's technique (BIANCHI; PIVA, 2012b), original and forged regions are discriminated in double compressed images, either aligned (A-DJPG) or nonaligned (NA-DJPG). Yang et al. (YANG et al., 2014) propose an error-based statistical feature extraction scheme to face the challenging case where both compressions are based on the same quantization matrix.

In most cases the analyst can exploit coding traces to disclose non-native images or to localize the tampering, reaching FD1 and FD2 in the forensic scale; FD3 has not been deeply investigated but, as shown in literature, coding traces can reveal something more than mere localization of the tamper. Farid (FARID, 2009a) shows that, when combining two images with different JPEG compression quality, it may be possible to recover information of the original compression quality of the tampered region. This technique has been proved effective only if the tamper was initially compressed at a lower quality than the rest of the image; on the contrary, when the compression is stronger in the latter

stage, the traces of the first compression are probably damaged and the detection fail.

Multiple operations and consecutive compressions can undermine coding traces, specially stronger compression later in the processing chain. The analyst using these tools should take into account the reliability of the results based on the coding characteristics of the investigated image (FERRARA et al., 2015). Chu et. al (CHU; CHEN; LIU, 2016b) tackle this issue proposing a framework to evaluate the detectability of multiple operations. This framework is validated by testing the effects of resizing, contrast adjustment, blurring, and double jpeg compression.

3.2.3 Editing Traces

Image editing modifies the visual information of the image and the scene depicted, introducing traces in several domains of the image such as pixel, geometric, and physical. Editing traces are the most numerous, and can be split into subcategories (Fig. 3.2) according to these domains.

Image illumination inconsistencies (light source direction, cast and attached shadows) are powerful traces considering that it is hard to achieve a perfect illumination match when composing two images. The are two main approaches for illumination techniques: geometric and illuminant. The first one is based on the geometric constraints of light, trying to use scene elements as cues to determine if the arrangement of lights (KEE; FARID, 2010) (CARVALHO; FARID; KEE, 2015) or shadows (KEE; O'BRIEN; FARID, 2014)(PENG et al., 2017) are plausible. Illuminant techniques exploit the color, intensity and temperature aspects of the illumination, and are able to detect if a region or object in the image was lighted by a different type of light (RIESS; ANGELOPOULOU, 2010) (CARVALHO et al., 2016).

Similarly, geometric relations within an image (e.g., object proportions, reflections) are determined by the projection of the 3D real scene onto the image plane. This process is commonly modelled through the pin hole camera model (HARTLEY; ZISSERMAN, 2004). Any deviation from this model can be exploited as evidence of tampering. Iuliani et al. (IULIANI; FABBRI; PIVA, 2015) uses a perspective constrained method to compare the height ratio between two objects in an image captured under general perspective conditions. Without the knowledge of any prior camera parameter, it is possible to estimate the relative height of objects and eventually identify objects that have been inserted in the scene. Conotter (CONOTTER; BOATO; FARID, 2010) describes a tech-

nique for detecting if a text on a sign or billboard has been digitally inserted in the image. The method looks if the text shape satisfies the expected geometric distortion due to the perspective projection of a planar surface. The authors show that, when the text is manipulated, it is unlikely to precisely satisfy this geometric mapping.

Specifically detecting copy-pasting (or copy-move) forgeries is a densely studied problem in the literature (ASGHAR; HABIB; HUSSAIN, 2016)(MAHMOUD; AL-RUKAB, 2016)Ṫhis type of forgery involves duplicating regions and performing operations such as stretching, rotation, and most forensics techniques focus on some sort of patch or region matching, exploiting locally non variant features (ZANDI; MAHMOUDI-AZNAVEH; TALEBPOUR, 2016), which can be considered editing traces.

When an editing trace exposes evidence of forgery, we can expect to infer something about its nature (FD3): if an object has as shadow inconsistent with the scene, he was probably inserted; if the illuminant color is inconsistent, the object could have been either spliced or retouched.

Obtaining other specific information about the techniques involved in the tampering process (FD4) is a very challenging task. There are two main reasons for this. The development of a technique for detecting the use of a specific tampering process/tool may require a strong effort compared to its applicability in a narrow range. Secondly, proprietary algorithms have undisclosed details about their implementation, making hard to develop analytical models for their traces. A first step toward this kind of assessment have been proposed by Zheng et al. (ZHENG et al., 2015) to identify the feather operation used to smooth the boundary of pasted objects.

3.3 Image Composition

Recent research on Image Composition were surveyed to determine which ones could be used to aid in forgery. For this purpose, techniques that a forger could use to perform any form of operation were considered, from splicing to highly creative operations. The techniques were classified in five general classes based on the type of forgery they could perform:

- **Object Transfering**: transfering an object or region from one image to another image, or even to the same image. This is the most common type of forgery, and encompasses both splicing and copy-and-paste operations. It is divided into *Alpha*

Matting, *Cut-Out*, and *Gradient Domain*;

- **Object Insertion and Manipulation**: inserting synthetic objects into an image or manipulating an existing object to change its properties. It is divided into *Object Insertion*, *Object Manipulation*, and *Hair*;

- **Lighting**: altering image aspects related to lights and lighting. It is divided into *Global Reillumination*, *Object Reillumination*, *Intrinsic Images*, *Reflections*, *Shadows*, and *Lens Flare*;

- **Erasing**: removing an object or region from the image and concealing it. It is divided into *Image Retargeting*, and *Inpainting*;

- **Image Enhancement and Tweaking**: this is the most general class of forgery, and is related to what is considered retouching in the forensics literature. It is divided into *Filtering*, *Image Morphing*, *Style Transfer*, *Recoloring*, *Perspective Manipulation*, and *Retouching*.

It must be noted that some of the surveyed techniques could be used to perform more than one type of forgery in the classification. Erasing, for instance, is often performed by copy-pasting regions of the image to conceal an object. In this sense, a technique under the *Object Transfering* classification can be also considered in the *Erasing* class.

In this Section, we discuss each of the different forgery classes and their relation to the forensic traces and techniques. Most information about the effect of composition on forensic traces comes from performed tests (see Section 3.4). The sub-classes are not explicitly divided in the next subsections, but are highlighted for readability and navigation. For the most relevant classes, we provide a brief general analysis from the forensics point-of-view.

3.3.1 Object Transferring

This class contains techniques that can be used with an end goal of transferring objects between images or in the same image.

3.3.1.1 ***Cut-Out***

A fundamental task of transferring an object or region is defining its boundaries, and techniques that can help making good contours are classified as *Cut-out* (MORTENSEN; BARRETT, 1995)(HUANG; ZHANG; ZHANG, 2011). These techniques do not change the content from the source or target images, they only help in selecting a pixel area.

Most techniques to detect splicing or copy-and-paste are well-suited against *Cut-out* forgeries, because the pixel content is unaltered. From a forensics point of view, well-defined boundaries in the transferred region reduce the amount of information being carried from the original image. This might alter some traces and affect the performance of techniques based on those traces (SUTTHIWAN et al., 2010). A bad cut can also be easy to note visually, without the use of additional tools.

3.3.1.2 ***Alpha Matting***

One of the main limitation of transferring objects by cut-and-paste is that transparency is ignored. Hair, thin fabrics, glass, and edges may contain a mix of colors from the foreground and background of the source image. This can cause visual artifacts on the resulting composition, and the presence of foreign colors that can be used for traces. *Alpha Matting* techniques can estimate the transparency of a region in the image, which can be used to better extract it from the source image, and then composite on the target image (Fig. 3e-h). The visual aspect is the most critical with the use of alpha matting for object transferring, as it blends colors in borders and transparent regions, making convincing forgeries. In most cases greater transparency is present only on a small part of the composition, such as borders. The majority of the composited area remains unaffected as a regular splicing.

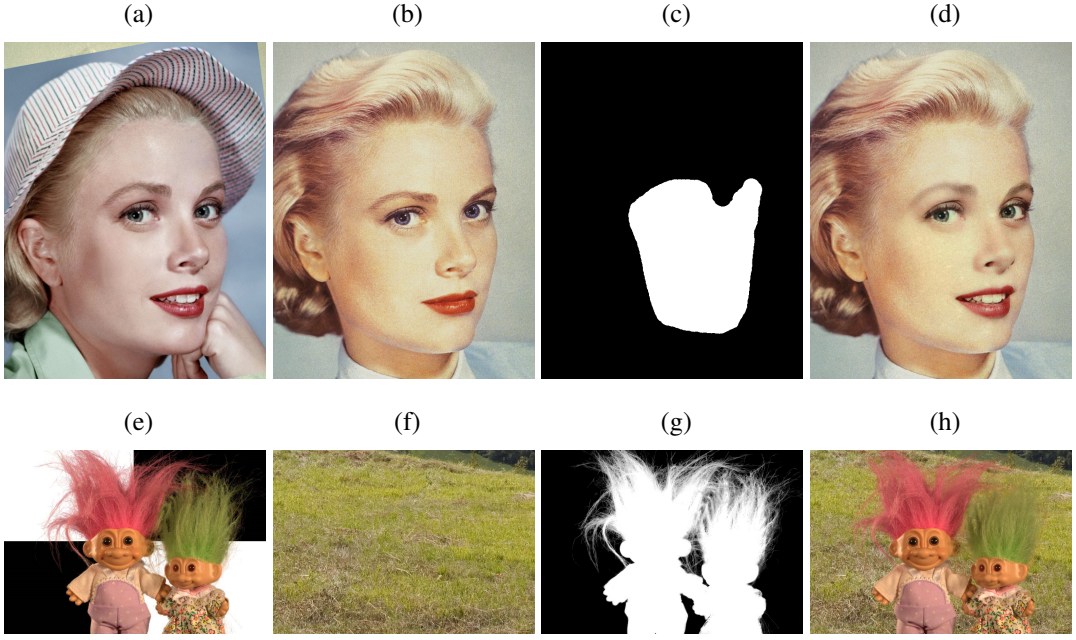
3.3.1.3 ***Gradient Domain***

The most sophisticated object transferring techniques are *Gradient Domain* ones. These techniques aim to combine the gradient of the transferred object with the target image, making a complex blend. The simplest technique is Poisson Image Editing (PÉREZ; GANGNET; BLAKE, 2003), which matches the gradients by solving a Poisson equation from the boundaries of the transferred region. The resulting object has different colors and gradient, blending with the scene. Poisson Image Editing, also commonly referred to as *Seamless Cloning*, spawned several papers that improved on its basic idea of solv-

ing differential equations for gradient matching of transferred regions (TAO; JOHNSON; PARIS, 2010)(DING; TONG, 2010).

Research such as Sunkavalli's (SUNKAVALLI et al., 2010) focus on the Laplacian Pyramid as the main component for sophisticated blends between images, being able to maintain the noise and texture of the target image (Figures 3.2e through 3.2h) to some degree. This kind of approach was generalized (FARBMAN; FATTAL; LISCHINSKI, 2011) and improved (DARABI et al., 2012) by other authors.

Figure 3.3: Example of splicing using object transferring techniques. The top row represents *Alpha Matting*, and uses the Shared Matting technique (GASTAL; OLIVEIRA, 2010). The bottom row corresponds to *Gradient Domain*, and uses Multi Scale Harmonization (SUNKAVALLI et al., 2010). The source images are in the first column, the target images in the second column, the transference masks are in the third column, and the final result is displayed in the fourth column for each technique.



Gradient Domain techniques can blend the whole transferred area and merge the images in a profound level. There are big variations on the inner workings of each technique, and the results are very dependent on the images to be combined. Furthermore, most of these techniques can be finely tuned. This makes them hard to be analysed from a forensics point of view. The safest way to detect forgeries of this kind would be focusing on high-level traces such as shadows and geometry. Light-based traces could help in cases where a full object is being transferred, because the resulting colors after the blending may create irregular lighting. When transferring parts of objects, such as changing faces in an existing head (Figure 3.2d), it is possible that the result can have plausible lighting and illuminant traces.

3.3.1.4 ***General Analysis***

Object Transferring techniques are arguably the most relevant to the forensics community, because they can be used to perform both splicing and copy-pasting. Figure 3.3 shows an *Alpha Matting* (top row), and a *Gradient Domain* (bottom row) splicing. Both forgeries are visually unnoticeable. Notice how the *alpha matte* (Figure 3.2g) contains very precise information about the transparency of each hair, and the mixture of colors in the final composition (Figure 3.2h). The *Gradient Domain* composition exemplified does not use transparency information (Figure 3.2c), but is able to transfer some of the color and texture of the target image (Figure 3.2b) into the transferred region of the source region (Figure 3.2a). The final result (Figure 3.2d) is a very convincing composition.

3.3.2 Object Insertion and Manipulation

Images are 2D projections of a 3D scene, with complex interactions of light and geometry. To insert a new object into the image, or to manipulate existing objects, the properties of the 3D scene must be known. This is a very challenging task. Techniques under this category focus on estimating characteristics of the 3D scene or its objects, providing means to alter them in a visually convincing way.

3.3.2.1 ***Object Insertion***

Rendering a synthetic object into an image is a simple task if the scene lighting and camera parameters are known. Additional knowledge about scene geometry also helps to increase realism. The challenge is to obtain this information from a single image. The most advanced techniques for object insertion, developed by Karsch, are able to estimate perspective, scene geometry, light sources and even occlusion between objects. In (KARSCH et al., 2011) heavy user input was needed to aid the parameter estimation, whereas in a second work (Figure 3.0b) (KARSCH et al., 2014) most input tasks were replaced with computer vision techniques to infer scene parameters.

3.3.2.2 ***Object Manipulation***

The manipulation of objects in an image suffers from similar problems than insertion. Scene lighting, camera parameters and geometry are required for a visually convincing composition, and the geometry of the object being modified must be also known. A slight advantage in relation to rendering synthetic objects is that the photographic texture of the modified object can be used, providing a more photo-realistic touch. It is possible to perform resizing operations on objects without directly dealing with its 3D geometry (WU et al., 2010), but most techniques will focus on modeling it.

The easiest way to work with the geometry of objects in an image is to limit the scope to simple primitives. Zheng (ZHENG et al., 2012) focus on cube-like objects, modeling them through “cuboid proxies", which allow for transformations such as scale, rotation, and translation in real time. Chen’s work (CHEN et al., 2013) uses user input to model an objects geometry through swipe operations. This technique works specially well on objects with some kind of symmetry, such as a candelabrum or a vase, and allows changes in the geometry itself. Another solution for dealing with object geometry is to use a database of 3D models, and find one that fits with the object depicted in the image (KHOLGADE et al., 2014).

Manipulating human body parts in images is a specially hard task, because human bodies vary greatly in shape, and clothes affect the geometry. This type of manipulation, however, is of special interest due to its applications in marketing photography and modeling. Zhou (ZHOU et al., 2010) uses a parametric model of the human body, and fits a photography to a warped 3D model, achieving a correspondence between body parts in the image and 3D geometry. This allows the reshaping of body parts, making a person in a picture look thinner, stronger, taller, etc. **Hair Manipulation** is also a hot topic in image composition, with a special focus on changing hair styles after the picture has been taken (CHAI et al., 2013)(WENG et al., 2013).

3.3.2.3 ***General Analysis***

Even though state-of-the-art techniques in image insertion and manipulation can create visually convincing results, they should not pose a problem for modern forensic techniques. Distinguishing between real and synthetic images is a very debated topic (FARID; BRAVO, 2012)(DANG-NGUYEN, 2014), and there are forensic techniques that focus on identifying them (PENG; ZHOU, 2014)(PENG; LI; LONG, 2013).

The weak point for this category of image composition is in the acquisition traces. The process of rendering a synthetic object is different from capturing it with a camera, so changes in the acquisition traces should point to the manipulation, providing FD1 or FD2 results. Similarly, when performing transformations on an object (scaling, rotating, deforming, etc.), its pixels have to be resampled, changing the acquisition traces. Resampling detection also could be used to obtain an FD3 result in these cases, while compression-based techniques could identify this type of manipulation if the original image was compressed. Kee has demonstrated that object insertion might be able to fool geometry-based lighting techniques (KEE; O'BRIEN; FARID, 2014), which could also extend to object manipulation. The reason for this is that the same lighting parameters estimated to verify the integrity of the scene were used to generate the composition.

3.3.3 Erasing

An erasing manipulation is when an element of the image is intentionally removed or hidden, and not a consequence of other editing. This category is comprised mostly of Inpanting and Image Retargeting techniques.

3.3.3.1 ***Inpaiting***

Inpainting techniques are used to complete a region in an image, filling it with appropriate content (BERTALMIO et al., 2000). By selecting a region that one wants erased as the region to be completed, inpainting can make objects disappear. Several papers are focused on stitching different parts of images together (HUANG et al., 2013) (KOPF et al., 2012), or filling large gaps (DAISY; TSCHUMPERLÉ; LÉZORAY, 2013). There are implementations of inpainting techniques already available on commercial editing sofware, such as Photoshop's Spot Healing Brush and Content Aware Fill tools. The main limitation of inpainting is filling regions with high amount of details, or using image features which are not local in the filling. Huang's (HUANG et al., 2014) work is capable of identifying global planar structures in the image, and uses "mid-level structural cues" to help the composition process.

3.3.3.2 ***Image Retargeting***

Image retargeting is a form of content-aware image resizing. It allows to rescale some elements in an image and not others, by carving seams in the image, i.e. removing non-aligned lines or columns of pixels (AVIDAN; SHAMIR, 2007). The seams usually follow an energy minimization, removing regions of “low-energy" from the image. The objects and regions that have seams removed will shrink, while the rest of the image will be preserved. This can be used to remove regions of the image by forcing the seams to pass through certain places instead of strictly following the energy minimization. Most research on image retargeting focus on better identifying regions in the image to be preserved, and choosing the optimal seam paths (PANOZZO; WEBER; SORKINE, 2012)(LIU; JIN; WU, 2010).

3.3.3.3 ***General Analysis***

Erasing manipulations should behave in a similar fashion to object insertion and manipulation, as the modified region will not come from a photograph, but from an estimation. This affects acquisition and compression traces, provided the original images were compressed. Image retargeting has already been analyzed from the point of view of image anonymization (DIRIK; SENCAR; MEMON, 2014), and there are papers focused on its detection (YIN et al., 2015)(WATTANACHOTE et al., 2015)(KE et al., 2016). Detecting that a seam carving has been done in an image would constitute and FD4 in our scale.

3.3.4 Lighting

Lighting techniques are capable of **changing the lighting of scenes** (WANAT; MANTIUK, 2014) and objects (LAFFONT et al., 2012)(BELL; BALA; SNAVELY, 2014), inserting light effects such as **reflections** (ENDO et al., 2012)(SINHA et al., 2012), **lens flare** (HULLIN et al., 2011)(LEE; EISEMANN, 2013), and even **manipulating shadows** (GUO; DAI; HOIEM, 2011)(FINLAYSON; HORDLEY; DREW, 2002). From a forensics point of view, lighting techniques are dangerous due to their potential of concealing other forgeries. After splicing an object in an image, for instance, a forger could add convincing shadows and change its lighting, making it harder for both human analysts and forensic techniques to detect it. Indeed, it is a concern in image composition when

the source and target lighting conditions are different, and there is research focused on correcting this issue (XUE et al., 2012a)(LOPEZ-MORENO et al., 2010).

3.3.4.1 ***General Analysis***

Due to the variety of lighting techniques, it is hard to make a general statement about them from a forensics point of view. As always, it seems plausible that at least an FD2 result can be achieved if compression is involved in the forgery. Techniques that add shadows or change the lighting in a visually convincing way, but do not account for all lighting parameters of the scene, could fail to deceive geometry and light-based forensics analysis. Specifically identifying light inconsistencies is an FD3 in our scale.

3.3.5 Image Enhancement/Tweaking

This is a broad classification for techniques that perform image modifications and are too specific to have their own category. **Image morphing** techniques (LIAO et al., 2014)(KAUFMANN et al., 2013) can fuse objects together, creating a composite that is a combination of them. **Style transfer** techniques are able to transform an image to match the style of another image (SHIH et al., 2014), a high-level description of a style (LAFFONT et al., 2014), or an image collection (HACOHEN et al., 2013)(LIU et al., 2014). In the same vein, **recoloring** techniques can add or change the color of image elements (CARROLL; RAMAMOORTHI; AGRAWALA, 2011), and even simulate a different photographic process (ECHEVARRIA et al., 2013).

Filtering techniques can be very flexible, allowing for a wide variety of effects. They can be used to remove noise or detail from images (Figure 3.0c) (GASTAL; OLIVEIRA, 2012)(CHO et al., 2014), or even to add detail (GASTAL; OLIVEIRA, 2015) while preserving edges. Different filters may be designed to obtain different effects. From a forensics point of view, filtering techniques can be used to remove low-level traces. A simple median or gaussian filter is able to remove compression and CFA traces, but it is easily detectable, as it softens edges. Edge-aware filtering, however, can be used to destroy such traces preserving edges. If used in a careful way, it can remove the aforementioned traces in a visually imperceptible way.

Perspective manipulation techniques allow an user to change the geometry (LIENG; TOMPKIN; KAUTZ, 2012), and perspective (CARROLL; AGARWALA; AGRAWALA,

2010) of a scene, or to recapture an image from a different view point (LEE; LUO; CHEN, 2011). Its uses are mostly artistic and aesthetic, but these techniques could be used to forge photographic evidence. The final type of manipulation that will be discussed is Retouching. Retouching techniques aim to perform adjusts on image properties such as white balance (BOYADZHIEV et al., 2012)(HSU et al., 2008), focus (TAO; MALIK; RAMAMOORTHI, 2013), or several at the same time (JOSHI et al., 2010). They can also aid in performing adjustments in several images at the same time (YÜCER et al., 2012).

3.4 Image Forensics vs. Image Composition

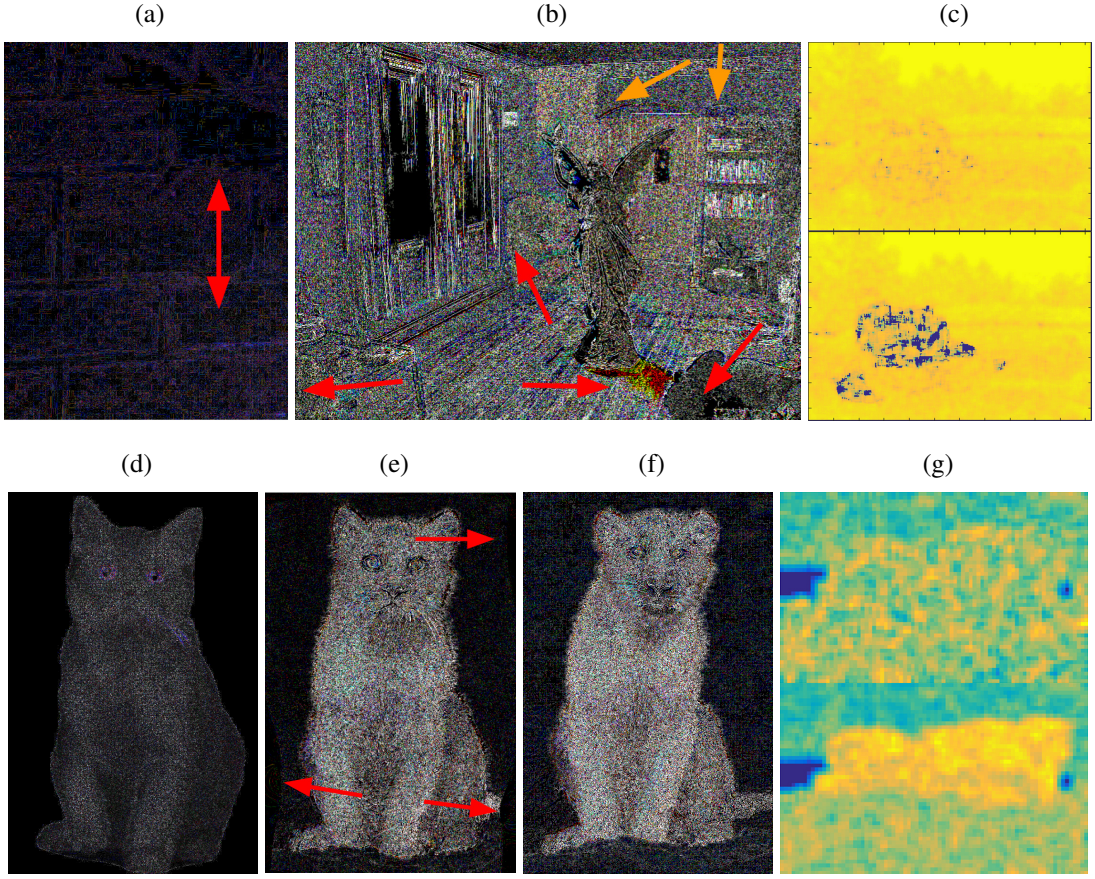
In addition to the general analysis in Section 3.3, to understand how image composition affects forensics traces, we performed a set of experiments. This task was challenging since there are no available implementations for most techniques. We separated our experiments into two phases: a more qualitative analysis considering a broad variety of techniques, and a more quantitative experiment focusing on a few state-of-the art techniques. The following two subsections discuss both phases in detail.

3.4.1 Qualitative Analysis

Firstly, to test on a broader scope how different image composition techniques affect forensics traces, we gathered images from 12 different publications on image composition, either from the publication website or directly from the authors. Approximately 80 images generated with 9 different types of forgery described in Section 3.3 were studied. Our main goal was to analyze the images directly before and after the composition has been applied.

In particular, we applied forensics techniques that analyze traces of CFA (FERRARA et al., 2012), PRNU (CHEN et al., 2008), Double JPEG compression (BIANCHI; PIVA, 2012b), ELA, and high-frequency noise. All these techniques generate as output a detection that can be used to visually identify if the composition had any outstanding impact on the corresponding traces. Strictly speaking, these techniques are FD2, but they may reach FD3 depending on what is uncovered. In Fig. 3.3b, the differently colored noise blocks suggest color overlay. Furthermore, the regular, square nature of some noise blocks indicate some sort of geometrical processing on the image. This is all information

Figure 3.4: Results of analyzing different traces for the images in Figure 3.1. (a) ELA of Soft Shadow Removal. In this case, it is not possible to identify any irregularity in the composited image. (b) Noise analysis of object insertion. The first identifiable irregularity is that the noise pattern for the shadow cast by the synthetic object greatly differs from other shadowed regions in the image (red arrows). The indirect illumination estimated after the scene's light interactions with the object appear as salient planes in the noise map (orange arrows). (c) PRNU analysis of localized recoloring. The more yellow, higher is the correlation between the region and the cameras sensor pattern noise. On the first image, there are some false positives thorough the image caused by high frequency areas. On the recolored image, the probability map shifts completely to the altered region. (d)-(f) Noise analysis of image morphing. The morphing process creates distinct warping artifacts on the noise pattern. (g) Double JPEG compression analysis of reillumination. The more yellow, higher the probability that the region has undergone double JPEG compression. While the top image shows a very noisy pattern, in the bottom image the uniform interpretation of a salient portion suggest that different compression traces (single and double) are present in the image.



that helps understand the nature of the alteration (FD3), even if in this case it depends on the expertise of the analyst. We reached the overall conclusion that most composition techniques affect low-level traces in some way.

Table 3.2: Overview of the tested scenarios, according to the level of detection for each trace. A green dot indicates the technique can be visibly detected by that trace (●), a blue plus is plausible (+), and a red x is visibly undetectable (×). Undetermined tested cases are marked as gray slashes (−), and missing symbols are non-appliable testing scenarios.

Composition Technique	CFA	D-JPG	ELA	Noise
Soft Shadow Removal (GRYKA; TERRY; BROSTOW, 2015)		−	×	●
Dehazing (FATTAL, 2014)	+		−	●
Object Insertion (KARSCH et al., 2014)	−	×	●	●
Reillumination (XUE et al., 2012a)		●	+	+
3D Object Manipulation (KHOLGADE et al., 2014)	+	−	×	×
Image Morphing (LIAO et al., 2014)	+	+	+	●
Alpha Matting (GASTAL; OLIVEIRA, 2010)(CHUANG et al., 2001)	●	+	+	×
Edge-Aware Filtering (GASTAL; OLIVEIRA, 2015)	●	×	+	●
Seamless Cloning (SUNKAVALLI et al., 2010)(FARBMAN et al., 2009)(TAO; JOHNSON; PARIS, 2010)	+	●	+	+

3.4.2 Quantitative Analysis based on JPEG Artifacts

For our quantitative experiment we focused on one of the most generally effective forensic approaches: image forgery localization via block grained analysis of JPEG artifacts, as proposed in (BIANCHI; PIVA, 2012a). This approach, by assuming that tampered images present a double JPEG compression, either aligned (ADJPG) or nonaligned (NADJPG), can be used to detect a suspect region. Once again, this is FD2, but in some cases FD3 can be reached allied with proper reasoning. If only a delimited region (such as a person or object) presents double compression, it is very likely that it has been spliced as-is from an already compressed image into an uncompressed image, and then compressed again. This is not common, as any adjustment done during the splicing operation, such as rotating and resizing would resample the image, destroying the traces of the first compression.

We replicated the experiments by considering the scenario where half of the image has undergone manipulation; but while in the original paper only splicing was considered, here we compared its performance considering three object transferring approaches: Splicing (SP), Alpha Composition (AC) (GASTAL; OLIVEIRA, 2010) and Seamless Cloning (SC) (SUNKAVALLI et al., 2010).

Similarly to (BIANCHI; PIVA, 2012a) we considered uncompressed TIFF images belonging to three different cameras (Nikon D90, Canon 5D, Lumix G2): 100 images were used for SP and AC while only 10 images for SC, due to its heavy computational cost (ten base images produced 1100 sample test images). They were acquired with the highest possible resolution and their central portion 1024×1024 was cropped. Then the following steps were performed for each image to produce A-DJPEG artefacts: i) JPEG compression with QF_1 was applied, ii) the left half of the image was replaced with the original TIFF applying each different object transferring technique, iii) JPEG compression with QF_2 was

applied. The NA-DJPG artifacts are produced by removing a random number of rows and columns between one and seven before step (ii).

The QF_1 and QF_2 are taken from the sets $[50, 55, \ldots, 95]$ and $[50, \ldots, 100]$ respectively. We performed our analysis using 6 DCT coefficients. The results were evaluated using the area under the ROC curve (AUC) by varying QF_2 (exactly as defined in (BIANCHI; PIVA, 2012a). In this way, we can aggregate many ROC curves in a single graph summarizing them by the AUC. AUC usually assumes values between 0.5 (random classification) and 1 (exact classification). In the following we discuss the achieved results, that are summarized in Figures 3.4a and 3.4b for the aligned and not-aligned cases respectively.

Alpha Composition: Since the result of the composition is strongly influenced by the value of α defining the transparency of the tampering pixel by pixel (see Section 3.3), to test all the possible outcomes we applied a linear transparency gradient mask from the bottom left to the upper right corner of the tampering, with four different α ranges: i) $[0, 1]$ - average response; ii) $[0, 0.3]$ - high transparency; iii) $(0.3, 0.7]$ - mid transparency; iv) $(0.7, 1]$ - low transparency. This means the bottom left corner had the lowest value, linearly increasing per-pixel until reaching the higher value in the upper right pixel of the image. The results confirm that both A-DJPEG and NA-DJPEG performance are strongly influenced by the α value: transparent objects can be hardly detected unless the last compression is really slight. Conversely, in case of low-transparency objects, there is no real difference between SP and AC. Considering that, in most real cases, high-transparency is applied only on a small percentage of the composition (like borders or hair), we expect that the use of this technique would not degrade the performance of the detection.

Seamless Cloning: The multi-scale technique allows to transfer the appearance of one image to another. It aims to to harmonize the visual appearance of images before blending them. Furthermore seamless boundary conditions are imposed to produce a highly realistic result. In order to exploit the peculiarity of this technique, the tampering region was slightly reduced, leaving a small border region out. The achieved results show that, similarly to the SP case, the detector produces an almost random output when the second compression is too strong. Anyway, when QF_2 is high, the detector is still able to detect the tampering, although with lower accuracy with respect to the SP case.

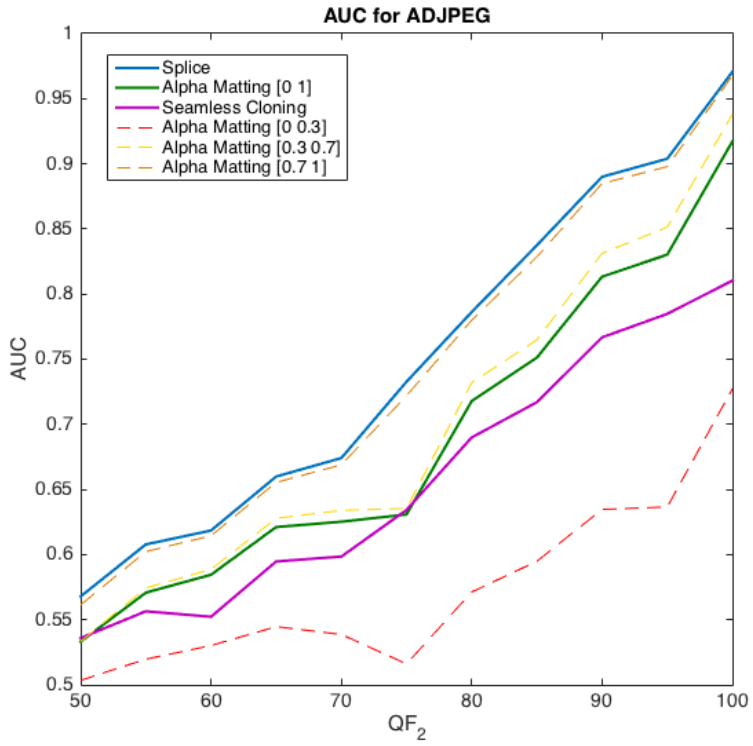
3.5 Summary

This chapter presented a survey of the fields of Image Composition and Digital Image Forensics, crossing them directly. A new classification scheme for techniques in both fields, along with a forgery detection scale were presented to help organize the discussion. This scale provides a clearer way to understand forensics scenarios and how the outputs of different techniques relate. To understand the forensics aspect of composition techniques, their inner workings were studied and tests were performed for a wide variety of image effects. Furthermore we assessed the applicability of an effective forensic technique for splicing detection against different kind of object transferring techniques quantifying how the performances may depend both on the kind of artifact (aligned or not-aligned double compression), and the parameters introduced by the composition technique (e.g., transparency factor in alpha composition).

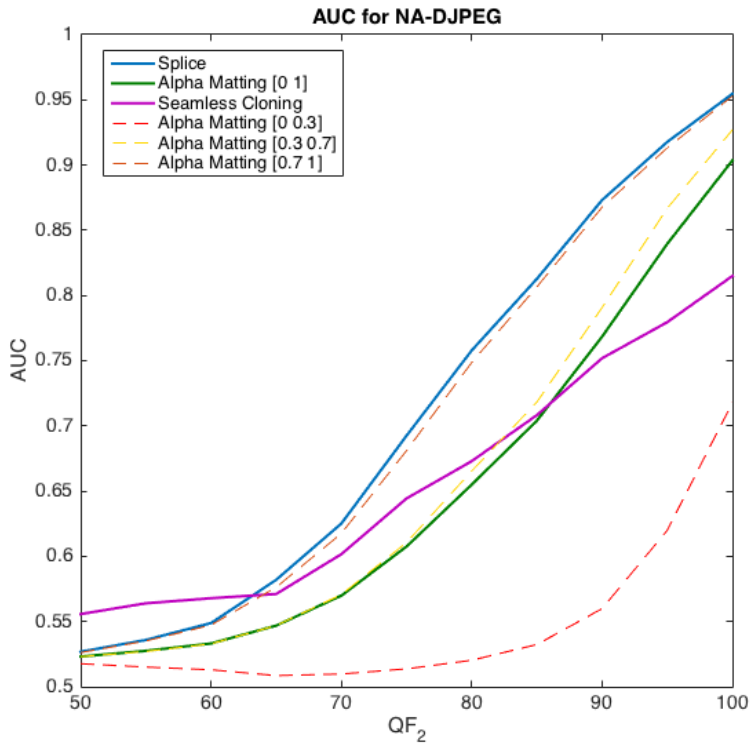
As a result, *we have shown that current state-of-the-art forensics has all the basic tools it needs to be able to detect most forgeries*, provided that they are properly tuned for each specific case, and maybe even combined.

Figure 3.5: (Best viewed in colours): Performance comparison of the A-DJPEG (3.4a), and NA-DJPEG (3.4b), against splicing, alpha composition and seamless cloning tampering. Dotted lines show the impact of tampering transparency on the performance. Note these are not ROC curves, but rather they show how the AUC varies when the second compression factor changes.

(a)



(b)



4 PLANNING FOR FORENSICS ANALYSIS

The results exposed in Chapter 3 demonstrate that most types of forgery can be detected using existing forensics tools. The challenge in analyzing an image then comes from choosing the appropriate techniques, understanding their outputs and articulating them. This chapter describes the development of a novel approach to tackle this problem: an automated planning system for DIF. Forensics techniques are described using a high-level language, focusing on the type of assessments and information they might uncover in an image. Given an objective for the analysis (*e.g.*, checking for double JPEG compression, forgery, etc.), this high-level description is translated into a logical equivalent and used to articulate a plan of analysis. Such an approach is based on the hypothesis that *forensics analysis tasks can be represented as a class of planning problems*. Here we demonstrate a prototype system that proves this hypothesis. The system is able to load a database of forensics knowledge described in XML format, and uses it to guide an image analysis process.

First, we introduce the concept of plans within the context of DIF, and use it to define the task of analyzing images as class of planning problems. Then, we explain the architecture of the proposed solution, its implementation strategies and practical results. The final section exposes challenges and limitations of automated planning systems.

4.1 Forensics Analysis as Plans

Since there is no definitive forensics tool or technique, it follows that the inspection of any image is a continuous process of analysis, comprised of many steps. This is intuitive, and in practice what many experts do, even if not explicitly. An example of inspection could include:

1. Visually inspecting the image for cues;
2. Inspecting meta-data such as format, size, header;
3. Executing technique A, and analyzing its result;
4. Executing technique B, analyzing its results and comparing with the results of technique 1;
5. Zooming in a suspect region of the image looking for further details;

6. Executing technique N, and so on, until
7. Concluding something based on the evidences (or the lack of them) from the previous steps.

By the end of this sequence of steps, a verdict is reached (*e.g.*, the image has been edited). In practice, forensics experts tend to use a limited set of techniques from the literature (IULIANI, 2016). In general, DIF experts are trained in toolboxes or analysis environments, which feature only a small amount of techniques. In this sense, the same person might always stick to a similar sequence of steps, or plan, for most analyses.

Having access to additional forensics techniques can increase the ability to uncover information, but also further complicates the analysis process. Ideally, one should be able combine, compare results from different techniques (*e.g.*, item 4 of the list) to reach a conclusion. Automatically combining results, also known as *decision fusion*, is a challenging problem as discussed in Chapter 3. Different techniques tend to produce different assessments about different things, with varying degrees of confidence and specificity (Section 3.1). The analyst is essential to this task of combining assessments, and our approach to assist him/her is based on automated planning.

The main goal of using automated planning for forensics analysis is providing the analyst with adaptable, coherent plans of analysis. Instead of having fixed steps, our approach uses all available information about the image at the time to determine what is the next best step. Since each step uncovers additional information about the image, the course of analysis may change. The specific knowledge about each technique, its interactions, etc, is contained within the system. In this way, the cognitive effort in memorizing many techniques, the results produced by them, and relating them at time of analysis is greatly reduced. Another advantage of such an approach is that it can help inexperienced users to perform professional-like analysis of images, given that the required specific knowledge is contained within the system. It also supports some forms of automation (for instance determining a sequence of non-assisted techniques to be run).

Automated planning is a field of artificial intelligence (NAU; GHALLAB; TRAVERSO, 2004) that deals with generating plans to solve problems. Tasks are formalized as state-machines that change with actions, and are solved by determining the sequence of steps (or transitions) between a starting and a goal state. A classic example is a warehouse with boxes and a forklift robot. The state-machine, also called the *state of the world* is the definition of all boxes' positions, plus the robot's position and orientation. The possible actions for the robot are turning, moving forward, picking up, and dropping a box. Given

the starting positions of boxes and the robot in a warehouse, and the desired, ordered state, the output is a sequence of actions to reach the desired state. For instance:

1. Turn right;
2. Move forward;
3. Lift box;
4. Turn left;
5. Drop box;
6. Turn left; and so on, until
7. the warehouse has been organized in the desired state.

Each action changes the state of the world atomically, and each step assumes this new state of the world, on a chain of steps leading from the starting state to the goal. To prove it was possible to treat the task of forensics analysis as a planning problem, our first challenge was to formalize it within this paradigm. Despite the similarities between the expert's list of steps, presented previously, and this type of list, transitioning between the two domains is not so straightforward.

4.1.1 The Planning Domain Description Language

In the warehouse problem, the state of the world is the set of all boxes' positions, and the forklift robot's position and orientation. The actions are what can be performed by the agent. Not all actions can (or should) be performed at all times, however. If there are no boxes in front of the forklift, the "Lift box" action has no use. Similarly, when in front of a box, it should not be possible to move forward. All these constraints have to be defined in the problem domain. Formally, a planning task can be defined as a 4-tuple in the form $\Pi = (V, A, s_0, s_*)$ (NAU; GHALLAB; TRAVERSO, 2004), where:

1. $V = \{v_1, ..., v_n\}$ is a set of **state variables**;
2. A is a set of *actions* a, where each a is a pair (pre_a, ef_a) of partial attributions; pre_a are the preconditions (of variables within V) for that action to be available; and ef_a their effects, *i.e.*, the changes on the variables of V;

3. s_0 describes the starting state of V, and s_* is a list of partial attributions for the final state, essentially the **goal**.

The Planning Domain Description Language (PDDL) (MCDERMOTT et al., 1998) was used to model our task as a planning problem. PDDL is based on a logical programming paradigm, and it describes variables as logical predicates. The predicate *position box1 x_0 y_0* could describe that there is a box at the position $(0, 0)$ in the warehouse. If at a certain time the state of the world V contained both this predicate, and *position forklift x_1 y_0*, it could indicate the agent is next to the box, and the action of lifting the box can be performed in *box1*.

PDDL conceptually separates a problem domain from each instance of the problem. If the rules and relations for solving forklift problems are defined, this is the problem domain, and it can be used to solve all instances of problems with varying starting states (s_0), and goals (s_*). The problem domain contains not only all possible actions, but also types of relations and possible predicates that are used in solving the problem. For instance, a description in PDDL of the lifting action could be the following:

```
;;;Lifts an adjacent box and holds it
        (:action lift_box
                :parameters (?b)
                :precondition (adjacent ?b)
                :effect (holding ?b)
                                (increase (total-cost) 1))
        )
```

Source Code 1: Example action in PDDL for a forklift robot.

The parameter for the action is a certain box *?b*, and its precondition is that it is adjacent to the agent. This relation has to be defined in the problem domain, and one of its definitions, as mentioned before, could be *(position box1 x_0 y_0) and (position forklift x_1 y_0)*. The question mark in this case expresses a parameter, a latent variable, instead of an instance (box1 is an instance of ?b). In its bare form, PDDL does not deal with numerical calculations, and most relations have to be explicitly defined, therefore another separate definition for adjacency would be *(position ?b x_0 y_0) and (position forklift x_0 y_1)*. The effect would be adding the predicate *holding ?b*, where *?b* is an instance of box, to the state of the world V. The expression *(increase (total-cost) 1)* is part of an optimization extension of PDDL, which allows for optimizing plans according to the costs of actions. This is discussed further in the text.

4.1.2 Dealing with Uncertainty and Non-Determinism

Lifting box1 (or any other box) changes the state of the world, so that *(position box1 x_0 y_0)* ceases to be true. There are a few different ways to model this fact, but essentially the predicate should be removed from V. This is achieved by adding the negation of the predicate (*i.e.*, *(not position box1 x_0 y_0)*) to the action effects. What this entails is that in PDDL all predicates are either true or false, and all true predicates are contained within the current state of the world, or are universal constants. The state of the world is fully known, and this does not translate well to the forensics context. Furthermore, the warehouse problem is completely deterministic; each action produces an expected output, and there is no chance for failing. In contrast, the steps taken by a forensics expert for inspecting an image are neither deterministic, nor provide full knowledge of the state of variables.

Classical planning essentially performs a graph search and, therefore, does not deal with non-determinism and uncertainty. Using the $\Pi = (V, A, s_0, s_*)$ definition, it expands a graph of possible states and finds a path between s_0, and s_*. In contrast, the outcome of each technique used by a forensics expert as steps in an inspection plan is not clear from the beginning. If it was, there would be no need for analysis. The task of inspecting an image is about *knowledge discovery*.

The solutions in the planning literature to deal with non-determinism are more complex to model, and their planning engines are not so straight-forward to use. Our solution to the problem is a hybrid approach, using classical planning in a flexible way. It can be summarized in a few core concepts:

1. Techniques and inspections produce information about an image, adding predicates to it. For instance, the predicate *(jpeg img)* indicates that *img* is a jpeg image;

2. If techniques and inspections change the state of the world by adding predicates about an image, they represent actions in the planning context;

3. The non-existence of a given predicate in V is treated by PDDL as the negation of that predicate. In practice, a complementary predicate is used to denote this knowledge. Thus, instead of evaluating *(jpeg img)*, one should instead use *(jpeg img) and (k_jpeg img)*, where *(k_jpeg img)* represents the knowledge about the property *jpeg*, be it true or false for *img*;

4. Since it is impossible to know the true outcome of an action before execution time,

no plan is definitive. The best outcome is expected for all actions, and in the case this is not true, a new plan is generated.

Items 3 and 4 deal, respectively, with the uncertainty and non-deterministic aspects of the forensics analysis task. If an action requires the image to have a particular property, for instance, a technique that only works on jpeg images, its preconditions will include *(jpeg img) and (k_jpeg img)*. Whenever the planner decides to use such technique in the plan, it first needs to execute an action that determines the truth value of *(jpeg img)* and adds the predicate *(k_jpeg img)* to the state of the world. If the image turns out not to be jpeg, the plan ceases to be adequate, and a new one is generated.

Figure 4.1: Illustration of the plan re-generation process in 4 steps.

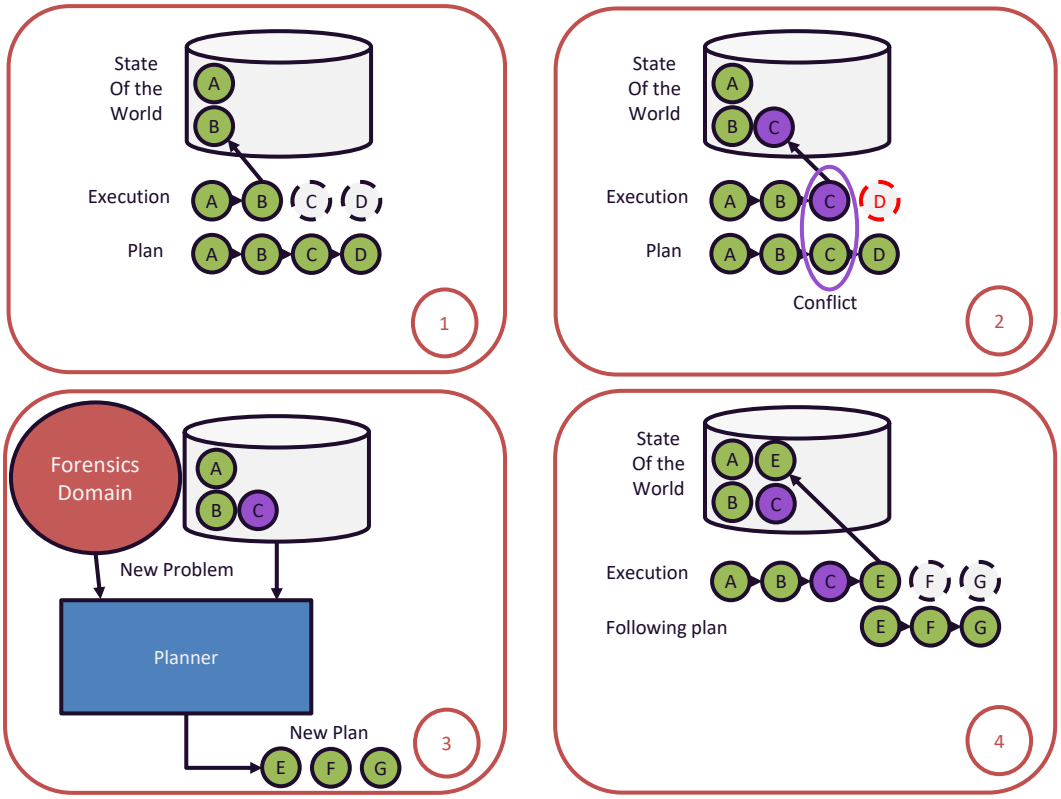
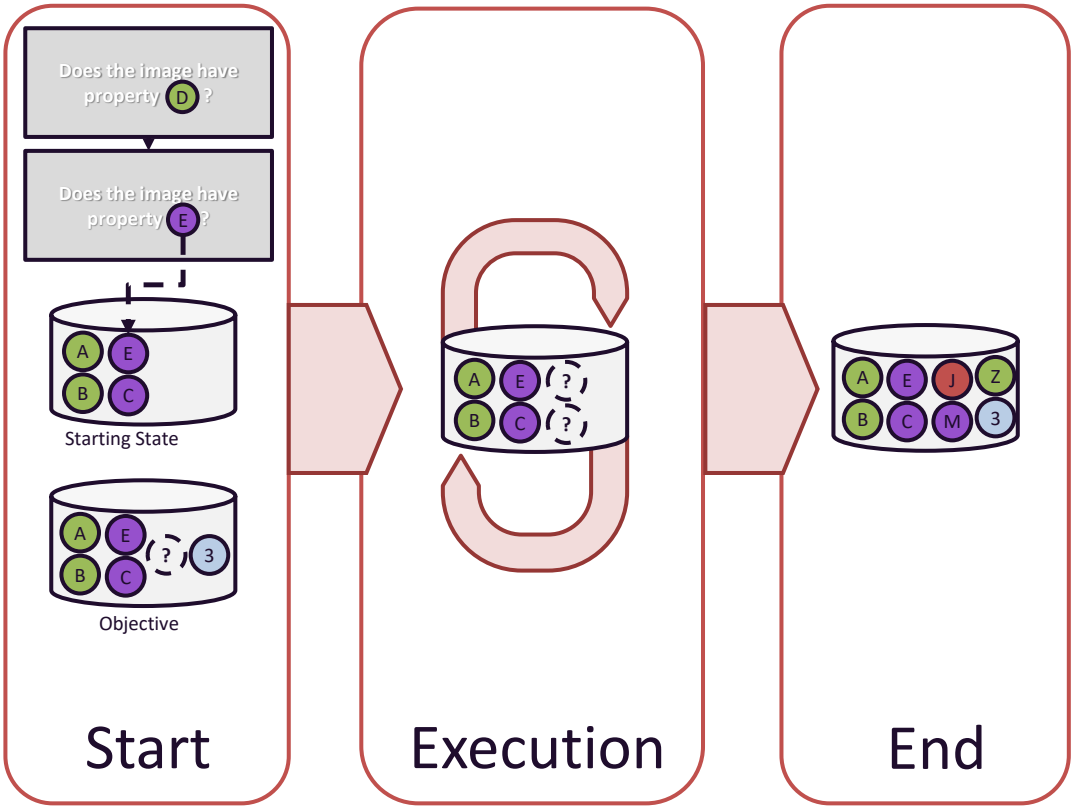


Figure 4.1 illustrates the plan re-generation process in four simple steps. Each colored circle represents a property or assessment about the image being inspected, associated with an action on the plan. A *green A* could mean, for instance, having the jpeg property, which is checked on the *A* step. On the top-left image (1), the plan expects to run *A*,*B*,*C*, and *D*, and receive the *green* output in all cases. Executed steps add their knowledge to the *State of the World*, regardless of the outcome. On the second image (2), executing *C* returned a *purple* output, instead of the expected *green* one. This conflict is detected, and it means the plan has to change. Nevertheless, the image possesses the

purple C property, and this is added to the *State of the World*. A new plan is generated by the planner, but this time the starting state of the world s_0 corresponds to the current *State of the World*. The objective s_* remains the same, and the set of actions and properties A remain the same, represented by the *Forensics Domain* used as input in the bottom-left image (3). The newly generated plan then follows executing its steps normally, accumulating knowledge about the image until the goal is reached.

Figure 4.2: The three phases of the analysis process.



The resulting flow of analysis can be seen in Figure 4.2. Any prior knowledge about the image can be added in the starting state s_0, thus optimizing the plan. This is illustrated using a series of questions, which is one of the ways a user could setup the analysis. The available information defines s_0. In the setup, an objective is defined as the goal state s_*. This is discussed in detail in the next subsection, but in Figure 4.2 this is represented by the *blue 3* circle, which is only one property, but it could be a set of goal properties s_g. The question mark circle indicates this is only a partially defined state, and any set that fulfills this condition is a valid final state. In other words, the execution process increasingly uncovers information, inflating the *State of the World* until $s_* \supset s_0 \cup s_g$ is achieved. The final phase (End) contains all the knowledge about the image obtained in this analysis. In this case, in addition to $s_{*0} \cup s_g$, the properties *red J*,

purple M, and *green Z* have been uncovered.

4.1.3 The Goal of Forensics Analysis

Until now the terms "goal", "objective", and "final state" have been used loosely, almost as synonyms. This helps understanding the general idea of planning in the context of forensics analysis, but there are important distinctions to be made. As previously stated, classical planning deals with deterministic scenarios, where s_0 and s_* formally define boundaries. In the warehouse example, the starting state s_0 corresponds to the physical setup of objects, and s_* to a different physical setup of the same objects. What is the actual *goal*, in this case? It could be many, for instance, organizing the warehouse. Imagine a fully automated workshop where robots are constantly working and building things, requiring materials from the warehouse. Sometimes, a truck comes in to be loaded with boxes. The same planning design can be used to solve many tasks in this scenario:

- When a robot requires certain materials in his workstation, a forklift robot needs to bring them to it. This task can be achieved setting s_0 as the current state of the warehouse, and s_* as any state for which the required box is at the robot's workstation. This task only cares about the final position of that one box, but many things can be moved in the process;

- After a full shift, the warehouse needs to be cleaned and re-organized, with every box in its right place. s_0 corresponds to the current state of the warehouse, but now s_* explicitly defines a place for everything, a fixed organization state s_o;

- When a truck comes to be loaded, its loading bay must be cleared of all boxes. s_0 is the current state of the warehouse, and s_* describes a state where no object is positioned in the loading bays, so the trucks can come in and park. If $s_* \supset s_0$, meaning there was no object in the loading bay to begin with nothing needs to be done.

In PDDL, the same domain can be used to solve all of these tasks, changing only the input problem. The high-level goal is explained first, and then its actual translation in terms of final state (s_*). The goal, the objective of the *planner* is to transition states until s_* is reached, but this is different than the goal of the *task*. The usefulness of any planning approach is limited by how well the goal of a task can be expressed as a final state s_* to be

reached. For the warehouse example this is very straightforward, but forensics analysis tasks are less well-defined.

As discussed in Chapter 3, the DIF literature lacks standards that could be used to properly define inspection tasks. Even a high-level description of an inspection task is open-ended, as there is no clear end. If there is a specific suspicion, and evidence is found to prove it right, then the inspection process can be finished. However, if there is no specific suspicion, or after repeated actions no evidence is uncovered, when should one stop? How is s_* defined in such cases?

The simplest scenario is when the goal of the inspection consists in knowing about a property explicitly modelled, such as *jpeg* (meaning the image is in the jpeg format), or *gamma_corrected* (if the image has undergone gamma correction). We model the knowledge about a property using the prefix *k_*, so if the objective of an inspection was to know if the image is jpeg, s_* is s_0 with the additional predicate *k_jpeg i*. The following snippet (Source Code 2) describes an action in PDDL that can be used to achieve such goal:

```
;;;Asks if the image is in jpeg format
        (:action ask_jpeg
                :parameters (?i)
                :precondition (not (k_jpeg ?i))
                :effect (and
                                (k_jpeg ?i)
                                (jpeg ?i)
                                (increase (total-cost) 1))
        )
```

Source Code 2: .

This action corresponds to asking a user (*e.g.*, using some interface) if the image is jpeg. It requires as a precondition that this is not known *(not (k_jpeg ?i))*. The description in PDDL of an automated technique that checks if the image is jpeg would be practically the same, since this is irrelevant for the planner. The key aspect for the planner is the state of *(k_jpeg i)* in terms of precondition and effect. As explained before, the planner always assumes the best outcome for a technique or atomic inspection. In this case, the effect is not only that we know about the property jpeg, but we know it to be true (*i.e.*, *(and (k_jpeg ?i) (jpeg ?i))*). The existence of the predicate *(k_jpeg i)* without *(jpeg i)* indicates we know about the property jpeg of the image, and it is not true (*i.e.*, the image is not jpeg). This would be a valid form of s_* if our objective was uncovering this property.

This simple example barely justifies the use of a planner, as the plan would have

only one step, executing the action *ask_jpeg*. However, from an user's perspective, it gives a clear way to define objectives for analysis. A user might not know the proper technique, or set of techniques needed to uncover a certain property, and in this sense the planer is performing a breadth-first search instead of a depth-first search (of subsequent steps). Furthermore, a property can be defined as set of other sub-properties or predicates, helping to encapsulate higher-level concepts.

In our experimental prototype, we make use of the FD scale (Section 3.1) to describe the property of an image being forged. Different techniques produce different types of assessments, or evidence from an image (FD1 to FD5), or none (FD0). We define a number of FD assessments from a set of techniques to collectively account for an image being forged. For instance, obtaining at least one FD3, and one FD2 output from different techniques. In this case, the goal is to set the property *(forged i)* as a result of obtaining one FD3 and one FD2 outputs. In essence, the final state s_* requires *(forged i)*. In practice, this is a continuous process of trying to prove the image has been forged, with the least number of steps possible. The solution continuously generates new plans, and runs techniques until either *(forged i)* is achieved, proving the image has been forged, or it has exhausted all tools.

```
;;;jGhosts technique
        (:action t_jGhosts
                :parameters (?i)
                :precondition (and
                        (not (run t_jGhosts ?i))
                        (k_jpeg ?i)
                        (jpeg ?i)
                        (roi ?i))
                :effect (and
                        (run t_jGhosts ?i)
                        (fd2 t_jGhosts ?i)
                        (afd2 t_jGhosts ?i)
                        (increase (total-cost) 4))
        )
```

Source Code 3: PDDL description of the jGhosts technique as an action.

Source Code 3 describes the PDDL code for implementing the described approach. A few interesting details should be noted. Firstly, under this strategy, actions that are inspections, such as techniques, should not be allowed to run more than once with the same parameters. This is required for the analysis process to evolve, otherwise it would be locked in an endless loop whenever a conflict arises, trying to re-run the conflicting step.

This is achieved by predicate *(run t_jGhosts ?i)*, which is present both in the precondition and effect of the action. It needs to be false (*i.e.*, this action has not been run with these parameters), and after being run the predicate is pushed into the state of the world.

The predicate *(roi ?i)* has an interesting role, denoting that the image has a region of interest. Some techniques require that the user selects a region, which usually contains a suspect object in the image. When analyzing an image, however, sometimes it is not clear if there is a forgery where it is located. If the planner decides a particular action is best suited to achieve the goal, and it requires that the user has a region of interest, it will generate steps to provide it. One way to achieve this is asking with a dialog if the user already has a region of interest. This can be part of the analysis setup (Fig. 4.2 "Start"), or obtained as the effect of another action. For instance, the output of a global forensics technique can outline a suspect region, providing a region of interest for further inspection. This is also closely related to our concept of *locality* in the FD scale, denoted by FD2.

The effects *(fd2 t_jGhosts ?i)*, and *(afd2 t_jGhosts ?i)* represent a *true positive* output for this particular inspection, steps to achieving our goal. Since native PDDL does not implement numerical operations such as addition, we use a token system to account for the accumulation of FD results. The first predicate (*fd2 t_jGhosts ?i*) means an FD2 result was obtained by the technique *t_jGhosts* on the image *?i*, and the second predicate (*afd2 t_jGhosts ?i*) generates an advancement token for the fd2 counter (hence *afd2*). Another action with cost zero is used to consume advancement tokens in order to progress through predicates indicating increasing FD assessments. For example, if the image had already one technique which successfully produced FD2, it should be *(fd2_1 i)*. Another successful FD2 technique would produce an advancement token that could be consumed by an action that would have the effects of *(not fd2_1 i)*, and *(fd2_2 i)*. This also means that all these predicates have to be explicitly defined in the planning domain (*i.e.*, having *(fd2_1 i)*, *(fd2_2 i)*, ..., *(fd2_n i)*, along with their transition state rules. In our first prototypes, this was hand-coded, but the forensics system described in the next section uses python to automatically generate virtual PDDL instead.

4.2 Architecture and Implementation

The core concepts to our planning approach to DIF have already been explained. Given a problem domain in PDDL with well-described forensics techniques, and an ob-

jective, the output will be a helpful sequence of steps. The feasibility of such solution lies in a meaningful PDDL representation of forensics knowledge, and its usefulness in the practicality of the process. To achieve this, we designed not only the back-end implementation aspects, but an architecture that allows a distributed maintenance of the knowledge domain. The general architecture of our solution is described in Fig. 4.3. It has four main components:

1. The **Forensics Domain**: The body of forensics knowledge that is used to generate sensible plans. Information about image properties, techniques, their behavior, and interactions is stored here in a high-level description language;

2. The **Database of Forensics Techniques**: It stores the actual implementations of techniques, which are executed when requested during analysis;

3. **Forensics System**: The core component, it coordinates the interaction with the user, maintains the internal state of analysis, uses the Forensics Domain to perform planning queries to the Planner, and calls the execution of techniques. Its internal components are seen on Fig. 4.4;

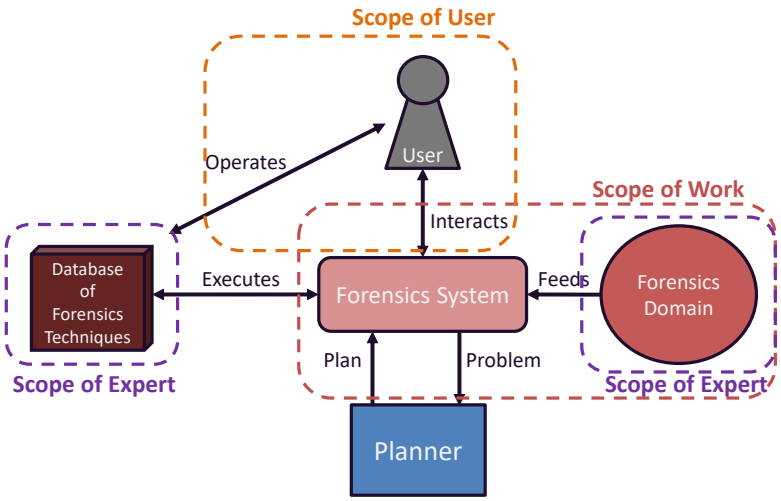
4. **Planner**: The planning engine, which receives planning queries and returns a plan.

The representation in Fig. 4.3 also illustrates the expected use and roles of different entities. Through interaction with the forensics system, the user provides an image and a goal to the Forensics System, which uses the Forensics Domain to generate PDDL queries to the Planner. The Database of Forensics Techniques, interfaced by the Forensics System, allows the user to execute the available techniques. The different scopes outline the collaborative, distributed aspect of such solution. Maintaining a continuously updated and relevant description of DIF knowledge should be a task for forensics experts. When implementing a new technique, the author also could provide its high-level description for the Forensics Domain, for instance. This would provide to any user the possibility of performing forensics analysis with the aid of plans.

The main component of our solution, the Forensics System, is itself comprised of four different components. These components and their interactions within the architecture are outlined in Fig. 4.4. They are:

1. **Interface Controller**: It is responsible for the GUI aspects, interacting with the user and controlling the flow of the process through its interactions with the Forensic Analysis Manager;

Figure 4.3: System architecture.



2. **Forensic Analysis Manager**: Is is the heart and brains of the operation. It maintains the State of the World and is the hub of communication between all different parts;

3. **Planning Language Engine**: It generates PDDL code to be used as input for the planner, and interprets the planner's outputs;

4. **Domain Language Interpreter**: It has an analogous role to the Planning Language Engine, interpreting the high-level description language of the Forensics Domain.

The State of the World linked to the Forensic Analysis Manager is not the exact same S from our strictly formal planning description, but it borrows the name from serving a similar purpose. The planner itself is not a part of our system. For this, we use the Fast Downward Planning System (HELMERT, 2006), a third-party , fully-implemented engine, and each PDDL query consisting of domain plus problem is independently run inside it to provide the output plan. After receiving a plan, each step is executed in sequence by the Forensic Analysis Manager, constantly checking for conflicts (Fig. 4.1), and maintaining a consistent representation of known information. Furthermore, all Forensics Knowledge, translated by the Domain Language Interpreter is maintained in an intermediate representation format within the State of the World (Fig. 4.5).

The Forensic Analysis Manager deals consistently with three levels of representation, as illustrated on Fig. 4.6. On the highest level, both python and our XML Domain

Figure 4.4: Inner structure of the Forensics System, and its connection to other components in the architecture.

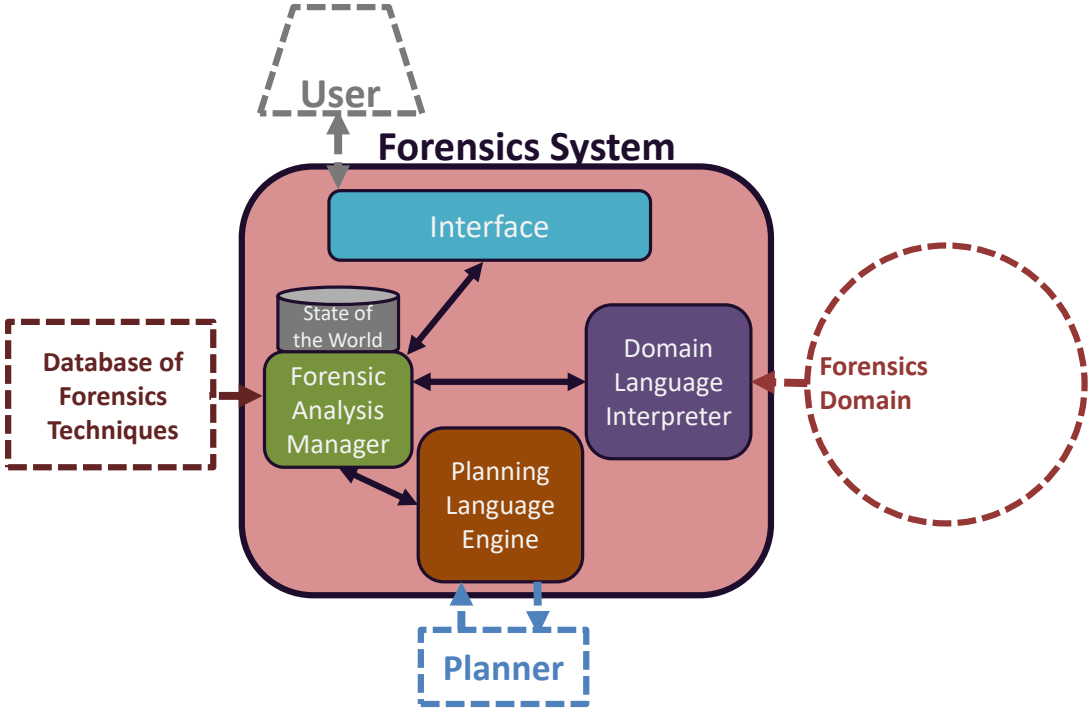
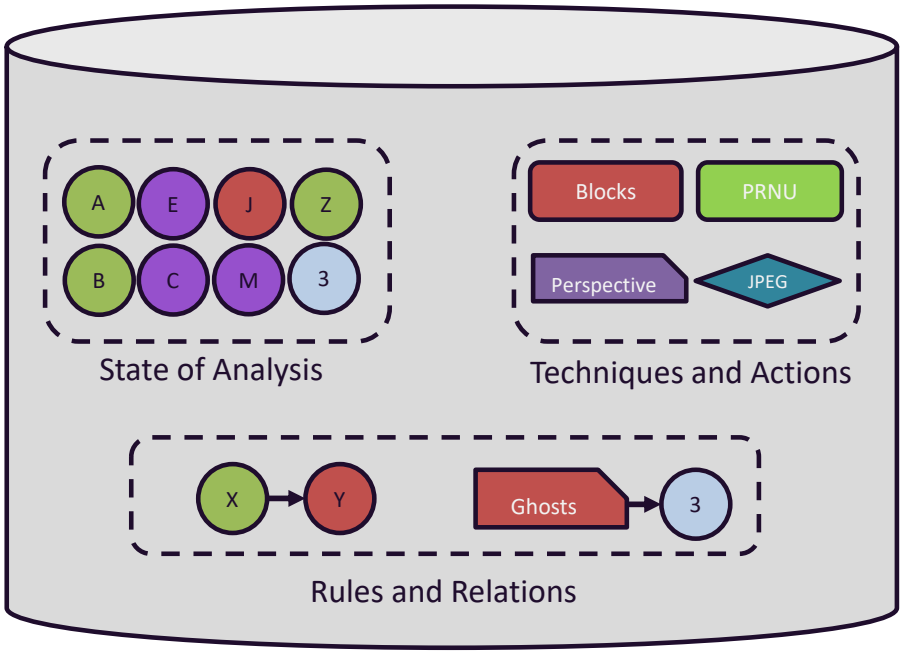


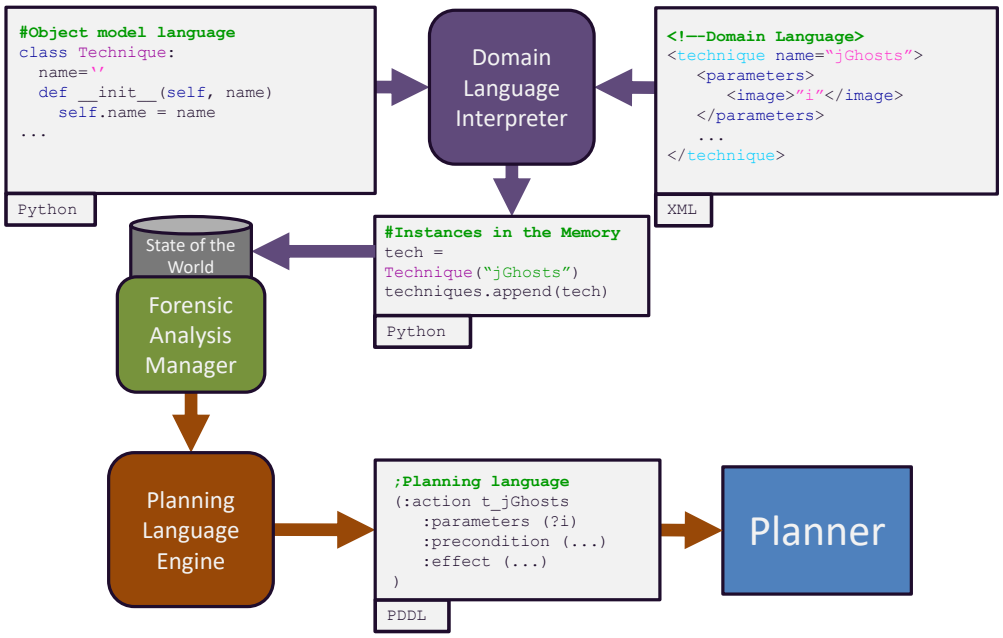
Figure 4.5: Internal representation of what is kept in the State of the World.



Language are used to describe the forensics techniques, properties, and relations among properties. The second level of representation are object instances in memory, realized when the Domain Language interpreter loads the Forensics Domain into the State of the World (Fig. 4.5). The third and lowest level of representation is PDDL dynamically gen-

erated by the Planning Language Engine, serving as input to the Planner.

Figure 4.6: The three different levels of representation of the forensics knowledge and its flow within the Forensic Analysis Manager.



A first PDDL prototype implementation proved it was possible to use perform automated planning for forensics analysis. the Forensic Analysis Manager allows flexibility between the low-level, rigid PDDL programming, and a high-level user friendly knowledge representation domain. To achieve this, we implemented a soft PDDL interpreter engine in python, focusing on the aspects needed for forensics analysis. This engine allowed us to maintain consistency between the logic used by the planner and our internal representation states. Since the plans need to evolve based on the updated based on the results of the executed techniques (Fig. 4.2), this was crucial.

The main features of our PDDL forensics engine are the manipulation of predicate lists and automatic generation of PDDL code. In PDDL, not only the state of the world S, but the different parts of an action's description (parameter, precondition, effect – see Source Code 3) are lists of predicates, either joint (and), or disjoint (or). Furthermore, there are abstract predicates, such as *(jpeg ?i)*, and their concrete versions *(jpeg img)*. We describe techniques as actions with lists of predicates for parameters, preconditions, and effects. Then, an algebra is defined to perform operation between those lists, for instance adding the effects of an action in the state of the world. Some predicate lists, however, can have different semantics, and this has to be taken into account. For instance, *(not jpeg img)* would deny an existing *(jpeg img)* predicate, effectively removing it from the State of the World.

4.2.1 The Domain Description Language

Our Domain Description Language, which is used to describe the Forensics Domain (Fig. 4.3) is based on the XML format. The main reason for this is simplicity. The burdensome PDDL notation is abstracted in favor of plain XML nested tags, and a few lines are sufficient to describe techniques and properties.

The description of a property in our format is seen on Source Code 4. The most important tags for properties are *name*, *type*, and *knowable*. The *name* must be unique, *type* indicates it is an image property (as opposed to a technique property), and *knowable* determines this property has to be discovered in some form, automatically creating an associated predicate (*k_jpeg*, in this case). The *analysis* tag allows to automatically create a question action for this property, *i.e.*, asking the user about the property. The associated PDDL code generated by this property description has many predicates plus the question action. The description part is not used directly for planning, but represents meta-information that can aid the analysis process. A user could learn more about a property or technique (s)he does not know about, but has appeared during his analysis. The more meta-information, such as links, associated publications, etc. is present in the Forensics Knowledge Domain, the more useful it is.

```
<property name="jpeg" type="image" knowable="true">
        <descript>"Jpeg compression format."</descript>
        <analysis="question"
        text="Is the image in the jpeg format?">
</property>
```

Source Code 4: Example of describing the "jpeg" property in the XML Domain Description Language.

Developing any sort of formal language, even if a purely description one, is a complex task, as design choices affect the limits of the language, *i.e.*, its representation power. The lack of planning can also cause unwanted consequences in how it is used, or purely in the "bug" sense. In our prototype, the Description Language was simplified as much as possible, and we limited the described entities to properties, actions, and relations, which were an experimental concept. The equivalent XML code for describing the *jGhosts* technique from Source Code 3 is presented in Source Code 5. It maintains a similar structure and elements from the PDDL version, with the additional meta tags for description, and type. The type in this case serves mostly as a categorization, relating to the discussed forensics technique types of Chapter 3.

```
<technique name="jGhosts" type="compression" cost="4">
        <descript>"This technique [...]"</descript>
        <parameters>
                <image>"i"</image>
        </parameters>
        <requirements>
                <hasproperty type="jpeg">"i"</hasproperty>
                <hasproperty type="roi">"i"</hasproperty>
        </requirements>
        <effects>
                <result type="fd2">
        </effects>
</technique>
```

Source Code 5: Equivalent XML description for the jGhost techniques presented in Source Code 3.

One of the most interesting possibilities for our solution is applying forensics knowledge that is not strictly a technique or inspection. Native, non-compressed image, for instance, should present CFA traces, or other form of acquisition traces. The absence of acquisition traces in a non-compressed image should hint for some sort of filtering, and compromise the nativity of the image (FD1). The "relation" tag is used to described this type of scenario in our Domain Language, and the equivalent description of this inference can be seen in Source Code 6. In practice, this does not generate PDDL code directly, but rather creates a relation in the State of the World (Fig. 4.5). Before executing a step in the plan, relations are checked, and, if valid, may add or remove predicates, triggering a re-generation of the plan.

The relation feature has great potential for practical use, but there are also different ways to implement it properly. It is unclear what the actual impact of separating relations from the actual planning process is. Checking for multiple relations and re-generating plans could prove to be cumbersome when the number of actions, properties, and relations grow. Other possible solutions were explored within the same architecture. For instance, implementing relations as actual actions, trying to force the required prerequisites to run it, as with actual techniques. Since we always assume an expected result, this creates bias and forces the planner on less meaningful paths, in the hope to prove the relations true. Another interesting approach is using the inferred properties not as actual discovered knowledge, but as some sort of "clue". The inferred filtering from the example could reduce the cost of running techniques that look for filtering traces, for instance. In practice, many ideas were discarded because they were too hard to test or validate in

```
<relation type="trace">
        <descript>"If the image is not [...]"</descript>
        <parameters>
                <image>"i"</image>
        </parameters>
        <requirements>
                <hasproperty type="!jpeg">
                "i"</hasproperty>
                <hasproperty type="!CFA">
                "i"</hasproperty>
        </requirements>
        <effects>
                <infer>
                        <hasproperty type="filtering">
                        "i"</hasproperty>
                </infer>
        </effects>
</relation>
```

Source Code 6: XML description of a relation to infer a property from other two properties.

realistic or useful scenarios. This is discussed in details in the next Section.

4.3 Results and Challenges

Our first and most relevant result in this front of research was showing that forensics analysis can be treated as a planning task. Given a proper description of techniques available and a defined goal, a sequence of steps can be generated to achieve it. One of such plans can be seen in Source Code 7. This plan assumes the goal is to obtain one FD3 result, evidence of the nature of manipulation that the image has undergone. An FD3 result is enough to put the nativity of the image in check, and serve as a starting point for further investigation, if needed. Additional evidence strengthens the case, so the goal can be changed after the plan is completed. What each step in Source Code 7 is doing is pretty simple:

1. Run an ELA (Error Level Analysis) on the image as a first exploring step. The appearance of a suspect region provides us with a ROI to be further analyzed;

2. Run a specific illuminant technique, such as Carvalho et. al.'s (CARVALHO et al., 2016), looking for inconsistencies among the illuminant color of the ROI and the

rest of the image. As explained in Section 4.1.3, the planner always assume the best outcome of the technique (FD3 in this case), and generates a token for advancing FD states;

3. The token is consumed by the state advancement action, indicating we have one technique with an FD3 outcome;

4. And our goal is reached. End of Analysis.

```
(t_ela img1)
(t_illumspec img1)
(adv_fd3_0 t_illumspec img1)
(goal_1fd3 img1)
; cost = 4 (general cost)
```

Source Code 7: Example of generated plan to perform forensic analysis of an image.

In our prototype, after a plan is generated each step is run individually. To describe the result of an action, the user selects from the set of its possible effects, determining which predicates are valid. An action to test if the image is jpeg will always add the predicate *(k_jpeg i)*, but *(jpeg i)* only if the result is positive, which is determined by the user. This allows us to test different scenarios without the need to actually interface with forensics techniques and their outputs. The generation of new plans is automatic once the State of the World reaches a conflict with the plan. For this, all the parts of the Forensics System (Fig. 4.4) were implemented. The interface is minimal, containing only the bare essentials for testing.

This setup allowed us to test all the required planning functions for our solution. It supports multiple independent goals, and composite goals that can be adjusted in real time. The user can also add his own predicates at any time, essentially injecting knowledge into the State of the World. If the injection does not create a conflict, the original plan can be maintained, but ideally this new information may be used to generate a better plan.

Our tests consisted on creating different Forensics Domain configurations, containing various techniques, and running simulations of analysis. The main limitation of such testing scenarios is that actual images and techniques are not part of it. For our purposes of showcasing the functionality of such an approach, instances of images and techniques were irrelevant. The initial goal was proving the consistency of such an approach, and its flexibility. Once this was established, the next step was a proper validation in practical cases, which was not full filled.

The first immediate challenge to validation was accessibility to the implementation of forensics techniques. The most straight-forward validation scenario would be to construct the Forensics Techniques Database along with its description in the Forensics Domain, and test it with users in actual images. However, the amount of scientific papers in forensics that provide implementations of their techniques is small. From the ones that do provide some sort of code or executable, there is a variety of programming languages, libraries and platforms used, sometimes incompatible. In our effort to gather a Forensics Techniques Database as large as possible we contacted authors, used virtual environments for compatibility purposes, and even tried to implement simpler techniques. In the end, we collected a few dozen techniques, some of which were just variations of a given method.

The techniques had different parameters and outputs, something which was expected in our original architecture (Fig. 4.3), and is abstracted in the "executes", and "operates" arrows. In practice, this requires a very tight interfacing between the Forensics Techniques Database and the Forensics System. All techniques should have wrappers, allowing interaction with assisted techniques, and the translation of their outputs as predicates to be inserted into the State of the World. Implementing this in an automated way has a strong impact on how the system behaves. For instance, if the output of a technique is a probability, a wrapper could be implemented that translated it into some *(property img)* if over a certain probability. What threshold should be chosen? 70%? 90%? This becomes, in essence, a parameter in our testing model. If the output of a technique is another image, the task becomes even more complicated, requiring some sort of algorithm to do the conversion. These issues were numerous, requiring a myriad of assumptions on external factors that, while not a part of our solution, had an impact on it.

To avoid a model crowded with ad-hoc parameters and derivatives of other authors implementations, eventually we pursued different validation strategies. Our goal was not to develop an actual software application, but rather explore the potential of our idea to tackle DIF problems. Two more approaches were studied for validation, one of them focused on user experiments with experts, and the other on automated tests.

4.3.1 User Validation of Plans

Instead of validating our proposed architecture, a user-centered validation was based on trying to show the usefulness of plans for forensics analysis. The natural extension being that if automated plans improved the process of analysis, our system was

one of such solutions to provide it. This would be tested by user experiments designed to mimic forensics analysis task with and without the aid of plans. For this end, we conducted interviews with forensics practitioners and experts to construct realistic scenarios. While it seemed clear for most interviewed professionals that such a system could be of use, it was hard to understand its place in their own workflows, or what sort of result they expected.

The draft test scenarios involved treating forensics analysis as sort of a puzzle for subjects. A set of edited and non-edited images, as well as available techniques would be chosen. The output of each technique for each image would be pre-processed, along with all possible plans for the analysis. During the test, the subjects would be presented the image, and be prompted to perform an analysis with and without the aid of the planner. When a technique was asked to run, the pre-processed result could be instantly given as output, registering the subject's choices, which would be analyzed later. This experiment would be performed both with inexperienced users and forensics experts. Our hypothesis is that being assisted by a planner not only reduced the cognitive effort in performing analysis, but provided a better analysis by suggesting ideal techniques for each situation.

From our previous experiences with tests on human subjects (Chapter 2), however, it was clear how our choices could bias the end results. The amount of techniques, images, which images would be chosen, even the pre-experiment explanation have to be carefully planned to control for bias. In general, using established methodologies, metrics, and data provides ground for the experiment. In this case, there are too many open variables and not enough material in the literature to help justify the choice of parameters. For a subject little or no experience with image forensics, it falls on the experiment design to brief him into this complex task, and then test it. For an expert, its hard to separate his or her degree of skill from his adaptation to the experimental setup.

The effort in designing a validation scenario abstracting implementation was very useful, even if it did not result in an experiment. It showed the difficulty in deriving meaningful metrics for DIF, specially outside a purely technical point of view. Furthermore, it questioned the use of instrumentation without a clear understanding the tools. In such scenarios, inexperienced subjects would be required to articulate judgment on images based on the output of techniques (s)he does not understand. If a correct answer comes from a wrong line of reasoning, how should this be treated? From our previous results (Section 2.5), people are almost as likely to ignore editing on images as they are to suspect something is wrong when it is not. This line of discussion led us to consider the

epistemological foundations of knowledge acquired through forensics techniques, and the possible Gettier cases (GETTIER, 1963). This is discussed in further detail in Chapter 6.

4.3.2 Automated Tests and DIF Set Theory

This subsection describes our attempts to perform automated, or non-assisted validation of our planning solution. Ultimately, we did not succeed in satisfactorily performing automated validation. However, our efforts to mathematically formalize some aspects of DIF provided important insights that are presented below. Our goal was to develop an abstract representation for DIF to avoid directly using images in our planning scenarios, allowing for broader experiments. Instead of having a jpeg image as a numerical entity with pixels that have undergone jpeg compression, the idea is to abstract it through an object with the "jpeg" property. This would avoid time and memory bottlenecks of actually processing images in experiments, focusing on the planning. For this to work, our abstract representations should be as formal and consistent as possible, a sort of "DIF Set Theory".

Fontani's (FONTANI et al., 2013) research on data fusion demonstrates the complex relation between different forensics techniques, and the phenomena they model. Some techniques have a positive synergy, in the sense that they corroborate on each other results, and some have a negative synergy, *i.e.*, a positive result on one should indicate a negative on the other. Understanding these relations is not only essential to combine them, but can be useful from a planning perspective. If we can assume that all pairs of techniques have either a positive, a negative, or no synergy between them, and that we can determine this relation, we can show the usefulness of the planner. At least conceptually, the planner can help to plan for synergism, and aid in the **discovery** of new synergies. For instance, *(synergy t_a t_b)→(synergy t_b t_a)*, but does *(synergy t_a t_b)* & *(synergy t_a t_c)→(synergy t_b t_c)*?

To test these sort of hypotheses, we planned automated tests running on abstract representations of images and forensics entities, such as properties and techniques. Different models for axiomatization were explored under Zermelo-Fraenkel set theory (ZFC, with the added axiom of choice) (CIESIELSKI, 1997). The main idea is that if a simple, conservative axiomatization of DIF that modeled some existing phenomena was achieved it could form the basis for a group theory of DIF, and explore morphisms.

ZFC is an axiomatic set theory that can be described with as little as eight axioms,

depending on variations. It starts with the empty set $\{\}$, or $\emptyset$, and through the successive application of axioms all other sets can be derived. Everything is a set under ZFC; there are no elements which are not sets, and no set can contain itself, thus avoiding Russel's paradox[1]. The natural numbers can be derived using von Neumann ordinals, in the following way: $0 = \emptyset$, the starting point for both the naturals and ZFC. Then, $n + 1 = n \cup \{n\}$, which gives:

$0 = \emptyset$

$1 = 0 + 1 = \emptyset \cup \{\emptyset\} = \{\emptyset\}$

$2 = 1 + 1 = \{\emptyset\} \cup \{\{\emptyset\}\} = \{\emptyset, \{\emptyset\}\}$

$3 = 2 + 1 = \{\emptyset, \{\emptyset\}\} \cup \{\{\emptyset, \{\emptyset\}\}\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, and so on.

While on the left side, black equations are simple arithmetic, the right side is derived from the application of the ZFC axioms. In this way, any countably infinite set can be defined, such as the natural numbers, or digital images. Digital images are a countably infinite set, even if we consider all possible resolutions, bit depths, and channels. For each resolution defined by a height h, width w, with a bit depth of d, and c channels, we can define the different images under a combination of parameters by $||I_{h,w,d,c}|| = h \times w \times d \times c$. Since all h,w,d, and c are natural numbers greater than zero, it follows $I_{h,w,d,c}$ is also a natural number greater than zero. Under this interpretation, a single black or white pixel can be represented by one bit, where $||I_{1,1,2,1}|| = 2$. The sum of all possible 2x2 black or white images of one channel is represented by $||I_{2,2,2,1}|| = 8$. $||I_{h,w,d,c}||$ grows fast when considering common digital images, with 32-bit depth and many megapixels, but nevertheless it is still a natural number. Since the addition of any two natural numbers results in another natural number greater then zero, it follows that the set of all digital images is infinitely countable, and can be mapped to ZFC ordinals.

There are a few different variations on how to perform this mapping, but it follows a similar logic to Cantor's diagonal argument. The idea is to define an order in how to increase the arguments h,w,d, and c, and list all possible images in them, in order. For instance, first the two images in $||I_{1,1,2,1}|| = 2$ are listed and paired with the numbers 0 and 1. Then we increase one in any dimension, such as $||I_{2,1,2,1}|| = 4$, and the four possible images will be mapped to the following numerals from 2 to 5. The mapping

[1] Let $R = \{x \mid x \notin x\}$, then $R \in R \iff R \notin R$. In other words, it is undecidable if a set containing all non-recursive sets contains itself or not.

from numerals to ZFC sets has already been shown, and under this construction all digital images can be mapped to ZFC, under a set I which is equivalent to the natural numbers. The main issue with this approach is the ambiguity in performing the mapping, since the 4-D manifold over h,w,d, and c must be fully covered. One way to achieve this is using Hilbert space-filling curves. For simplification, we can limit the problem to model a limited universe of images, such as $I_{800,600,8,3}$, the set of all 800x600 resolution, 3-channel images with 8 bits per channel.

In theory, this means that for a property p that an image may or not have, there exists a function P that returns a subset of I containing all images with that property. ZFC, however, does guarantee that for any property p there is a set of all things that satisfy that property implicitly. This is only true if this property can be described as a recursive construction of its subsets using first-order logic. This limits our ability to prove some things from this exercise, as describing image properties from the empty set using first-order logic is rather complicated.

The alternative we explored considers the ZFC mapping as latent space, a back-shadow that all images cast. We can describe properties of images in a more "high-level" and unconstrained formulation, and then obtain the translation to the ZFC ordinal domain. MATLAB (MATLAB, 2011), for instance, treats images as multi-dimensional matrices. Even if it allows floating point representation, pixels are either 32 or 64 bit discrete variables. In this way, an element of $I_{800,600,8,3}$ is an 800 by 600 by 3 matrix of 8-bit integers. With a very large (but still finite) amount of memory, one could have an 800 by 600 by 3 by 800x600x3x256 matrix with all possible images on $I_{800,600,8,3}$. For clarification purposes, let us call this "MATLAB-space" (I^M), which has an equivalent mapping to "ZFK-space" (I^Z) in our construction. Any testable Boolean function that could be implemented in MATLAB could ran on each member of MATLAB-space individually, and the indexes of the ones that return **true** could be turned into the ZFC equivalent. The union of the sets of the indexes with a **true** or **false** result by the function is the totality of $I^Z_{800,600,8,3} = I^M_{800,600,8,3}$.

Let us think of a simple property, $p_1(i)$, of images i where the sum of the intensity of all pixels on the red channel is larger than the green or blue channels. First, we take the sum of the intensity of all pixels in the red channel, and call R, then for the green channel G, and for the blue channel B. Images where $R > G \quad \& \quad R > B$ have this property, and even if it can only be properly defined in $I^M_{800,600,8,3}$, it projects into $I^Z_{800,600,8,3}$ by the mapping relation. A following property, p_2, where $R > (G + B)$ can be

defined, and it follows that $p_2 \rightarrow p_1$, so all images i in $I^M_{800,600,8,3}$ can be described as either $p_2(i)\ \&\ p_1(i)$, $\neg p_2(i)\ \&\ p_1(i)$, or $\neg p_1(i)$. In other words, $(p_2(i)\ \&\ p_1(i)) \cup (\neg p_2(i)\ \&\ p_1(i)) \cup (\neg p_1(i)) = I^M_{800,600,8,3} = I^Z_{800,600,8,3}$. This is a complete partition of I in terms of $p_1(i)$, as it appears in all terms.

Consider any forensics technique that can be reduced to a Boolean statement, and imagine it being applied to $I^M_{800,600,8,3}$. It can be seen as a function $T : I^M$ measuring a property $t(i)$, and again $t(i) \cup \neg t(i) = I^M_{800,600,8,3} = I^Z_{800,600,8,3}$. This is true for all deterministic techniques that can be reduced to a Boolean output, even if using a threshold. For instance, if a technique outputs a probability value (*i.e.*, a value $\in [0, 1]$), $t(i) \iff T(i) > 0.7$ would be a valid representation for a forensics technique. The motivation in this line of reasoning is to approach the DIF problem as a Boolean satisfiability problem (SAT), and derive implications. Since our planning solution already uses SAT-like formulations for its goals, this could provide for an interesting grounding validation.

A key observation is that even though $p_1(i)$ and $t(i)$ can be seen as properties that subdivide the image space, they are defined very differently. The property $p_1(i)$ can be determined simply by the content of the image, with no additional information needed, while $t(i)$ is described with respect to T. There is no ambiguity in what it means for an image to have the property $p_1(i)$, at least in a mathematical description. It could be roughly translated to English as "images that are predominantly red", or "images that have a red channel with greater intensity compared to the green and blue channels". In opposition, consider T as a technique that determines if an image is jpeg, for instance as an action to evaluate *(jpeg i)* in our planning solution. We can determine for practical purposes that $t(i) \iff$ *(jpeg i)*, but this is clearly untrue for many reasons. What does it *mean* for an image to be jpeg? It is not defined by any forensics technique T.

For simplicity, we are ignoring metadata and file extensions, and considering only content-wise what a property such as jpeg means. Both metadata and file extension are finitely countable, and they could be added to our I^M and I^Z definitions. Jpeg compression is an operation that transforms the content of an image according to some parameters, such as the quantization table and compression quality. Ideally, a jpeg image should be defined as an image that has undergone jpeg compression, but this definition is based on the history of the image, rather than on its content.

The way the definition of a *jpeg* property is generally approached is through an statistical modelling of the behavior of jpeg compressed images. In other words, statis-

tically differentiate the domain (I), and co-domain (I') of the jpeg compression function $J : I- > I'$. For some parameters of J, and for some i, however, $J(i) = i$, which generates ambiguity, and we cannot decide *(jpeg i)*. This can happen, for instance, if the quality factor is near 100%. What stands from this is that $I \neq I' \iff I \setminus I' = \neg j(i)$: we can only determine the *(not jpeg i)*, and only if there is a difference in the domain and co-domain. We can be sure when an image has **not** undergone jpeg compression under certain parameters, but for all the intersection $I \cap I'$ it is impossible to know for sure. The image could have undergone an identity jpeg transformation ($J(i) = i$), or only *looks* like a jpeg. This invokes the discussion from Fig. 1.1 on the Introduction: what is the difference between a photograph of an ocean, and an image that looks like an ocean? An important subject of research in this sense is the detectability operations (CHU; CHEN; LIU, 2016a), which is has an impact on anti-forensics (SINGH; SINGH, 2017).

Most authors recognize such limitations of their techniques, and disclose ambiguous or failure scenarios uncovered in their model. The practical approach is to determine a safe margin for parameters, based on tests and the literature. In terms of detection theory, if the $j(i)$ property is desired to be modeled, a technique $T_j : I$ tries to represent it through a virtual property $t_j(i)$. On a perfect scenario $t_j(i) \iff j(i)$, but in reality there are four possibilities:

1. $t_j(i)$ & $j(i)$: A true positive, when the technique is detected as jpeg and **we know** it is;

2. $\neg t_j(i)$ & $\neg j(i)$: A true negative, when the technique is not detected as jpeg and **we know** it is not;

3. $t_j(i)$ & $\neg j(i)$: A false positive, when the technique is detected as jpeg and **we know** it is not;

4. $\neg t_j(i)$ & $j(i)$: A false negative, when the technique is not detected as jpeg and **we know** it is.

The whole domain can be subdivided in one of these four sets, *i.e.*, $I = (t_j(i) \ \& \ j(i)) \cup (\neg t_j(i) \ \& \ \neg j(i)) \cup (\neg t_j(i) \ \& \ j(i)) \cup (t_j(i) \ \& \ \neg j(i))$. Notice that the term "we know" has been highlighted, as it is an important distinction. Any of those cases require actual knowledge of $j(i)$, which can only be achieved in test scenarios. If $j(i)$ was observable, there would be no need for T_j. The problem is that in reality T_j is never tested over I, or over any limited domain ($I_{800,600,8,3}$, for instance). Firstly, it

is computationally impractical, as the amount of possibilities is astronomical. Secondly, a large portion of the possible images in I is not meaningful for most purposes, being random noise or unnoticeable variations between one another. We can further define a subset of *real-world* images I_R, and a testing subset I_T, such that $I_T \subset I_R \subset I$.

All techniques are tested on a subset I_T, which contains members for various instances of I, such as $I_{800,600,8,3}$, $I_{1024,768,8,3}$, and so on. By constructing I_T, one can select images knowingly or forcibly containing the desired properties. When considering the universe of all possible properties and traces that can be scrutinized forensically (Section 3.2), it is impossible to predict how well I_T is representative of I_R. In Section 3.4 we presented both qualitative and quantitative experiments on different forensics techniques. Our results revealed unexpected behavior of some techniques tested with different I_T compated to the authors' original tests. While some techniques have proven to be very effective in detecting image features that they were not designed to detect, this raises concerns on the semantics of forensics analysis. Is it good or bad that a technique for splicing detection also detects other things? How can we decide between a true positive, false positive, or unexpected detection in such cases? Without the answer to these two questions, we were not able to implement a satisfactory validation experiment, and turned our attention to them.

4.4 Summary

This Chapter described a novel approach to forensics analysis, through the use of automated plans. It discussed both the theory and architecture behind our solution, and demonstrated a working prototype. The complexity in validating our solution comes from the many involved topics. There is the human element, present not only in the operational aspects, such as usability, but on the subjective nature of images. There are the logic and the linguistics aspects of describing forensics knowledge and the desired intentions for an analysis. In the technical side of DIF the main limitations are in the access and usability of current technology.

In "IQA: Visual Question Answering in Interactive Environments" (Gordon et al., 2017), the authors introduce the task of answering visual questions by autonomous agents. Given a high-level question about an environment, such as "how many mugs are there in the room?", or "is there a tomato in the fridge?", an agent interacts with it to provide an answer. There is a discovery process: walking around the room, visually inspecting

things, opening doors. In their architecture, there are high-level controllers, and low-level controllers. The core element of the higher-level controller is a planner, which outlines a discovery strategy to be executed by the low-level controllers, which would involve, for instance, navigation and object manipulation. An internal memory (analogous to our State of the World, Fig. 4.5) holds discovered information, and there is a constant reasoning using this information to formulate an answer. The remarkable similarity between their solution and ours stems from the fact that a forensics analysis is, essentially, visual question answering, too.

The authors describe the effort in designing test scenarios for their work, which is validated fully by simulation. Since the low-level operations being executed by their agents are mostly mechanic and well defined, such as walking to a certain position, it is easier to measure success. It is not sure how an adaptation of such an approach would perform to answer DIF questions, but hopefully this line of research can aid in providing a general framework for visual question answering.

5 THE LIMITS OF TECHNOLOGY

Chapter 2 has highlighted the need for technical solutions in analyzing images. On Chapter 4 we demonstrate the challenge of using the current state-of-the-art techniques discussed in Chapter 3. This Chapter builds a more general perspective of DIF, oriented by my vision of its practical role regulating information currency. We argue that DIF is constrained, firstly, by language and communication. As a regulator of trust in communication, it cannot over-extend its boundaries and become a limiter of communication. This could happen from a prescriptivist perspective, where the rules of DIF dictate the rules of visual communication with digital images, or from misuse. Therefore, DIF is secondly limited by its epistemic foundations. Incorrectly interpreting assessments and analysis could result from either mistake or malice, both of which depend on stretching the implications of forensics knowledge. On a system of distributed trust, all parties involved should be able to understand the basic rules, which is why epistemology is so important for DIF.

5.1 DIF Knowledge vs. Applications

If one is to look for a common goal or object of study in DIF publications, a picture starts to form, but it is not clear. This was discussed in the Introduction (Section 1.2). Intuitively, DIF aims to study digital images, providing knowledge and tools to analyze them, but in practice it produces artifacts and statements about different classes of things. Research on forgery detection, for instance, mainly produces techniques to detect inconsistencies on images, while camera identification tries to link an image to a capture device. On their experimental settings, however, both of them try to extract truth statements from empirical analysis of digital images. For instance, has this image been modified? True or False. Was this picture taken with this camera? True or false.

Many motivations have been used to justify the need for DIF and its various applications, but objectively one cannot avoid to ask: what is the practical use of all this? It is easy to come up many real-world use(r)s:

1. A judge needs to determine if a photograph can be trusted as evidence in his court;
2. A company needs to validate digitized documents against fraud;
3. A journalist wants to investigate if a publicized picture has been tampered to avoid

spreading misinformation;

4. A social media user wants to know if his friend is editing his pictures to make him appear more attractive.

Most researchers on DIF would agree these are desired and valid applications for their work, or that their research aims to aid in those cases. In practice, however, there are many gaps that prevent these practical uses. To understand why, let us instantiate each case:

1. After consulting a forensics expert, the judge decided against accepting the picture as evidence;

2. In a week, 12 out of 100 documents processed by the company were determined to be fraudulent;

3. The journalist found no evidence to suspect of a picture showing Trump and Kim Jong Un shaking hands;

4. After going through all of his friend's posted pictures, it appears he has edited some of them to look taller and thinner.

All of these can be simplified as truth statements with different questions:

1. Is the evidence adequate? False;

2. Are the documents frauds? 12 True, and 88 False;

3. Do we have reasonable doubt to suspect the picture? False;

4. Has my friend been editing his pictures to look more fit? True.

Being able to extract truth statements is an important part to formalize problems in the field, and in practice, it is the core of all used metrics. ROC tables, accuracy, and classification statistics as discussed on Chapter 3 are based on true positives, true negatives, false positive, and false negatives. A true positive classification depends on two truth statements: one about the class a sample was identified with (let us call it F(x)), and one about its actual class (C(x)). On a forgery detection context, it is usual to design experiments separating the images in two groups: images containing forgery (C(x) = True), and images not containing forgery (C(x) = False). In our study with subjects, we had to adapt the possible answers to these type of statements (Table. 2.1).

When a forensics technique is reported to have 90% accuracy, it means that over all tested samples, the proportion of true positives and true negatives over all classifications was 0.9. Sensitivity is the number of true positives over the number of true positives plus false negatives, and so on. It is clear, then, that the instrumentation of DFI is based on the evaluation of truth statements, both F(x), and C(x), and its derivations such as F(x) & C(x), where & is the logical "and" operator. The fact that both real-case applications as well as the scientific methodology of DIF are based on the evaluation of truth statements would seem to indicate that a framework for the formal representation of DIF problems is straightforward. Our efforts in developing a plan-based approach to DIF show this is not true. Digital images are images, used by millions of people to store information and communicate with different intents and outcomes. Being able to derive truth statements about digital images means to be able to derive truth statements about all these agents, actions, and systems. If we recall our previous examples, none of them are about the images themselves, but rather their use and implications:

1. The judge only cares about the image as it contributes to the narrative of the ongoing case. The image being pristine or not might be irrelevant if there is no suspicion of ill-will. If there is a jury, however, it is hard to determine the effect of questionable evidence in the court case;

2. Depending on the type of business the company is doing, the impact of fraudulent documents may vary. If they are simple registration forms and identification, having an automated system could deter most amateurs and discourage them from attempting fraud. If the company is victim of a large fraud, having this system could also be favorable to negotiate with their insurance provider;

3. For a journalist investigating a groundbreaking image, any outcome might be newsworthy. In the current days of fake news and click-baiting, there is enough audience and and there are divided opinions to capitalize on almost anything. Perform any sort of principled analysis on highly-publicized material can potentially draw a lot of attention;

4. It is very likely that the social network user knows his friend personally. If one is in doubt of the actual physical shape of a friend, the easiest way to clear this doubt is to meet him in person. The implication of why the images have been changed is what would be gossip-worthy.

To arrive at useful conclusions from those cases, there are many layers of complexity over the measured F(x) & C(x). If the individuals involved in them were to draw on the body of research in DIF, many assumptions would need to be made in the process. Each of them would add their own areas of expertise to derive their truth statements, be it legal (for the judge), fiscal (for the company), geopolitical (for the journalist), or social (for the user). If we extend this to all possible applications of DIF over all fields, there are many evident gaps that need to be addressed.

5.2 The Players and their Roles

The four cases that have been discussed were chosen because they represent plausible, easy to understand scenarios that could employ DIF. The exercise of exhaustively analyzing them from different points of view is fundamental, as it exposes questions that are not generally discussed in the literature. This is understandable, as any scientific work must limit its scope for practical research purposes. It is the interest of external agents, such as the market or society, to pick up parts and pieces and build real-life applications from research.

Let us assume DIF as a body of knowledge and techniques that can be easily accessed and drawn upon. In each of the cases, who is drawing upon this body of knowledge, and how is it being used?

1. In a court of law, usually the judge will bring in an expert. The judge has access to DIF through the expert;

2. In the case of a company that needs to validate documents against fraud, there are many ways this could happen. The company could have a dedicated department, develop its own technical solutions, or rely on consultants. For this analysis, let us assume the company uses an automatic, commercial software “S”. The company itself has no access to DIF, but rather instrumentalizes it. The developers that make the software access DIF;

3. As investigating images becomes a common practice in journalism, it is expected that journalists have access to DIF, or develop their own methodologies. A big newspaper could have dedicated forensics experts as well, but let us assume the journalist is investigating the image alone in this case;

4. A social media user investigating his friend's pictures will probably look in the Internet for literature or free DIF tools. This case is similar to the journalist on #3, but more likely with less resources and commitment involved.

Figure 5.1: Representation of the dynamics between agents in the four example cases. The cubes represent the object of analysis (images or documents). The green elements represent the agents that are interested in the result of the analysis, the brown lines represent the access to DIF, the red arrows represent the access to the object of analysis, and the analysis itself. The yellow agents are intermediates that have access to DIF, but are not the actual interested parties. The purple "S" represents the software used by the company to validate digital documents. The red outline on case 2 highlights that

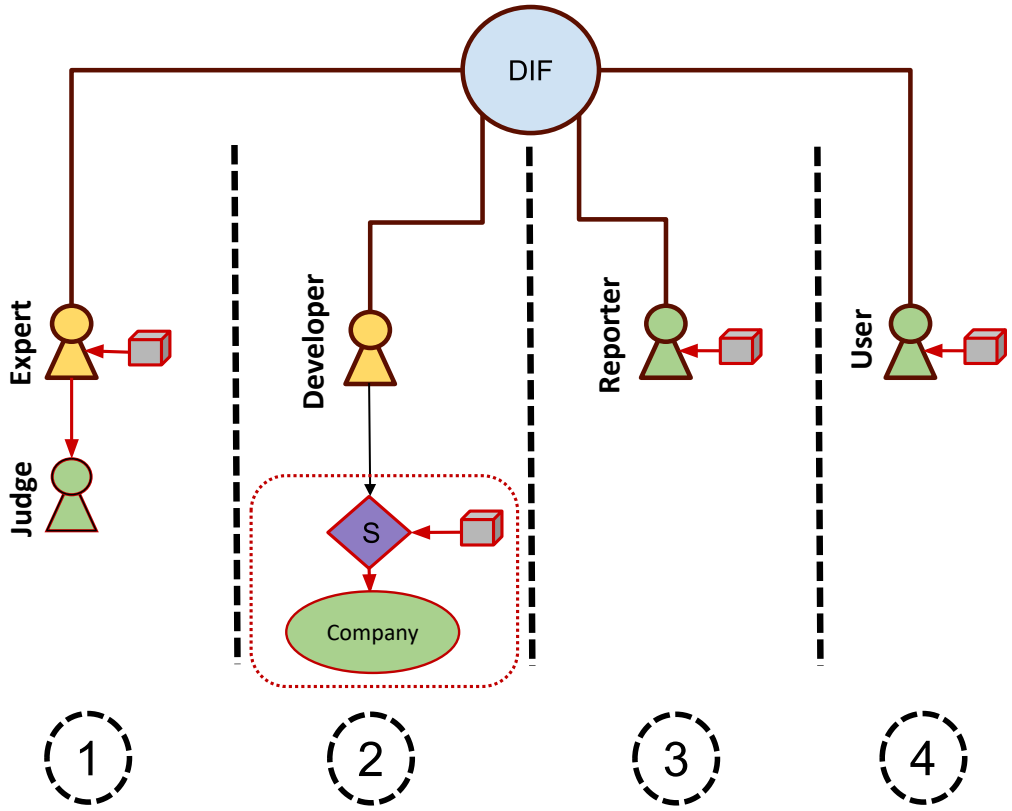


Figure 5.1 illustrates the dynamics among the involved agents in each case. In cases 3 and 4, the interested parties access both DIF and the object of analysis - a suspect photograph and a set of social media pictures, respectively. In this context, *access* is used to mean both *having access to* and *to access*. For instance, the judge from case 1 has access to the object of analysis if he wants to. Since he recognizes his limitations on understanding it, he trusts the handling to the expert. In case 2, the person that accesses DIF is the developer of the software "S". He does not access or have access to the object of analysis, *i.e.*, (s)he will not be handling the images during the analysis, which is done automatically by the software "S". Since the documents are property of another

company, (s)he also does not have access to them. If the company that uses the software had problems with some documents, the developer could have access to them for reasons of support and bug fixing, for instance.

One way to interpret this diagram is as the distribution of work, interest, and knowledge among the involved parties. Agents in brown connected to DIF have knowledge, agents connected to the red flow will do the work of analysis, but the actual interested in the result is in green. For cases 3 and 4, all the roles are concentrated on the same agent (the reporter and the social media user, respectively). In this analysis, then, they are equivalent, while cases 1 and 2 have a different structure. The judge from case 1 is the interested party, but delegates the work to the expert, who has knowledge. In case 2, "S" does all the work for the interest of the company without having access to DIF itself.

To understand the importance of this diagram, we must consider the stakes involved in each example (this is also discussed in Chapter 6. On all four cases DIF is needed with different degrees of urgency, and a mistake has different consequences:

1. A trial has serious consequences regardless of the case in question. Pictures can be used as evidence for all sorts of cases, depending on the particular legislation. In this sense, mistakes can not only result in wrong convictions, but jeopardize future cases by jurisprudence. A decision on a speeding ticket case with pictures from speed detectors, for instance, could create precedents for the applications of DIF in courts;

2. Depending on the kind of business done by the company, accepting fake documents can have a different impact, from financial losses to serious legal problems. In this example, the company is processing hundreds of documents per month, so it is likely that the net cost of a mistake is estimated and accounted for;

3. It is said that a picture is worth more than a thousand words, and we have many examples in history of pictures playing a key role in important events. It is hard to determine the effect of a controversial picture nowadays, which could range from public outrage to a drop in the stock market;

4. While we would like to believe that, on a social level, edited images have little impact, and in turn, DIF would be harmless, cyberbullying, harassing and doxxing are a reality. For the sake of this example, let us assume no ill-will from the social media user and limit the impact of analysis to the sphere of gossip.

5.3 Quantification and Situational Problems

One might be inclined to quantify the difference in stakes or the impact for each example, and the previous description even suggests an underlying descending order: 1, 2, 3, 4. Legal consequences and jurisprudence seem more important than the profit of a company, which in turn seems more concrete than the investigations of a journalist, and so on. This is not the point of this analysis, as I do not wish to instantiate each case beyond necessary to show my points. These are plausible examples that leave room for the imagination and background of the reader to fill in the gaps without getting in the way. That being said, trying to quantify exposes some of the issues that separate the DIF research and practical use. Consider the following (hypothetical) questions related to each example:

1. There are three experts available, one charges $100 an hour for his work, the other $1,000 for a complete analysis, and the third one $2,000 if he can find inconsistencies in the image, nothing otherwise. How should the judge decide which one to call in each case?

2. The software "S", has two problems: i) it says highly-compressed images are forged, even if they contain no forgery, which is a problem with customers; and ii) it does not detect a few types of changes in a document that a human would. Both cases could be solved by hiring an employee to inspect the documents. Problem (i) should be checked before "S" runs, to avoid discarding valid documents, and problem (ii) should be checked after, looking for special cases. Is it worth to hire employees to check the documents? Should they do the check before or after "S" ran?

3. The reporter got first hand on a shocking picture that could be used to write a groundbreaking story. If he writes and publishes a story on it as-is, his article would be the first and probably receive the most exposition. If the picture turns out to be a fake, he could risk fines and his reputation. He has some knowledge in DIF and some tools, but estimates it would take at least 4 hours to do a decent evaluation. By then, other reporters would have had access to the picture and he could lose his advantage, but if the image turns out to be fake he could have an edge. There is still room for mistakes on his side, and the only way to be sure is going through the forensics experts on the newspaper, which could take days. What should he do?

4. After going through social media pictures of his friend for the past 12 months, two seemed suspect, and one of them was posted by another friend. If he confronted his friend about the pictures, he could excuse himself easily. Going further back into older pictures could bring more examples, but the farther back the least interesting of a gossip subject it becomes. Should this person confront his friend, keep inspecting for older pictures, or just assume it cannot be known?

All these questions add more data to each case, but it never seems to be enough. In an effort to *quantify the unquantifiable*, we run into more problems. A good answer for all of them seems to be "it depends". There is a combination of variables such as time, risk, resources, trust, and quality of the analysis. All these variables could be instantiated with different weights complementing the relations diagram, and one would arrive at different answers for each question. One of the main motivations for the development of our planning approach was the flexibility to tackle these types of problems. Our solution could be part of a larger framework using optimization to maximize utility and minimize risk in a variety of scenarios. In this Chapter we show, however, that *this problem should not be tackled purely from a technical perspective*.

5.4 The DIF Oracle

I now invite the reader to explore, through a thought experiment, the limits of what is possible, plausible, and probable to be achieved with DIF under various circumstances. We can think of this as an asymptotic analysis of DIF to determine its bounds. Let us begin by considering two different worlds, where the main difference is the level of advancement of the field of DIF:

- World 1: DIF contains all possible knowledge about images at all times, and is ultimately infallible. There is an infinite amount of techniques, and they are all deterministic;
- World 2 (like our world): DIF has a limited knowledge about images, but we do not know how much of all possible knowledge it covers. There is a large amount of techniques, which might fail on some cases and/or be stochastic.

On both worlds, the limits of the analysis carried out by agents are bounded by the limits of DIF itself. In practice, however, it is also bounded by the capacity of the agents to

understand and use the knowledge of DIF, and translate it to real situations. This is what will be demonstrated. Let us now take a look at two variations of world 1:

- World 1.0: every agent that accesses DIF is ultimately limited by how much knowledge and experience he/she has and can use to solve problems;
- World 1.1: there are agents who can fully understand the knowledge of DIF. For all purposes they function as a DIF oracle.

In all senses World 1.1 represents the DIF utopia. The DIF oracle is the pinnacle of evolution for planning solutions, with full Forensics Domain (Fig 4.2), Database of Forensics Techniques, and integrated with an advanced AI system (such as IBM's Watson). What can we consider, then, to be reasonable questions to ask to the DIF oracle? If one has an image, presents it to the oracle, and asks questions about it, what can (s)he expect as answers? For instance, let us assume the judge from example 1 has access to the oracle, and decided to use it as an expert. A hypothetical court transcript of the conversation between the judge and the oracle follows:

- J: DIF Oracle, have you access to infinite knowledge about images, and computing power to evaluate images without making any mistakes?

- O: Yes, your honor.

- J: Can image 1 be considered a valid piece of evidence in this court?

- O: I do not know what qualifies an image to be considered a valid piece of evidence in this court, your honor.

- J: Well, is this image real or fake?

- O: What do you consider to be a real or a fake image, your honor?

- J: Does this image represent a real event? Were all the elements present in the image as they were represented, at the time the picture was taken? That would be a real image. Therefore, anything different would be a fake image.

- O: Your honor, all images are projections, they compress information into a much smaller representation form. They are not themselves representations of reality, but measures of the different interactions of light. The same scene can be captured

by infinitely different angles, with different capture devices and parameters. Conversely, there is an infinite amount of scenes that when photographed would result in the same image. There is no way to discern between a real place and a sufficiently reconstructed studio. Furthermore, all photographs have an exposure time, which is the amount of time it captures light from the scene. In this sense, it is integrating light from an infinite amount of small moments, not a single moment in time. To realize your definition of "real" for an image, one must have access to a massive amount of information, a great deal of which is from the past. In other words, to truly determine if an image is "real", one must have access to the past. I suggest a simpler solution is to ask the person who took the picture. Have you asked him?

- J: Yes, he is present in court, and has vowed the image to be real, but we cannot trust solely his answer.

- O: Why not, your honor?

- J: Because one of the people present in the photograph says it is not true.

The pedantic DIF oracle is not wrong in any of its answers, but does not seem to help the judge in any practical way. Since our assumptions include infinite knowledge of DIF and never being wrong, it follows the problem lies outside the knowledge of DIF. If there is an outside, it means there is an inside, which means there is a border between what can be considered knowledge in DIF, and what cannot. This is illustrated by Fig. 1.2 and Fig. 1.3. By the answers provided by the oracle, one could reinterpret Fig. 1.3 as shown in Fig. 5.2, arriving at a centralized, nested, or even self-containing scope. Even with access to complete knowledge and infallibility on those topics, the judge was not able to solve his DIF problem. To understand why, we model the knowledge/information relations as shown in Figure 5.3, and discuss them. The present actors are as follows:

- The DIF oracle: in this court, (s)he is the consulting expert. We assume (s)he has infallible knowledge about DIF, access to the picture, and we must assume some understanding of English, as he was talking with the judge;

- The Judge: the person conducting the trial. (S)He also has knowledge of English, access to the picture, and limited (probably not infallible either) access to legal knowledge;

Figure 5.2: Interpretation of the boundaries of DIF knowledge, based on the DIF oracle answers and assumptions of infallibility.

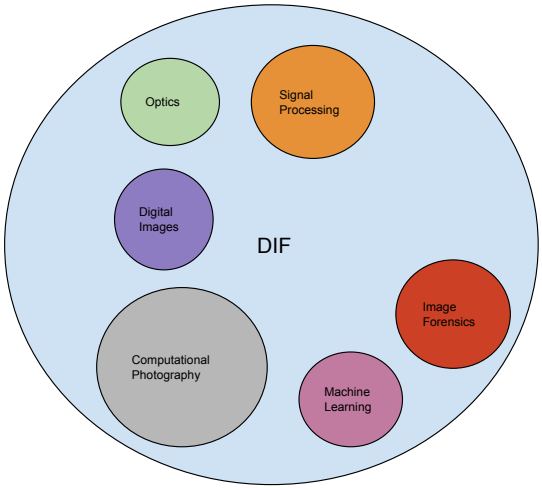
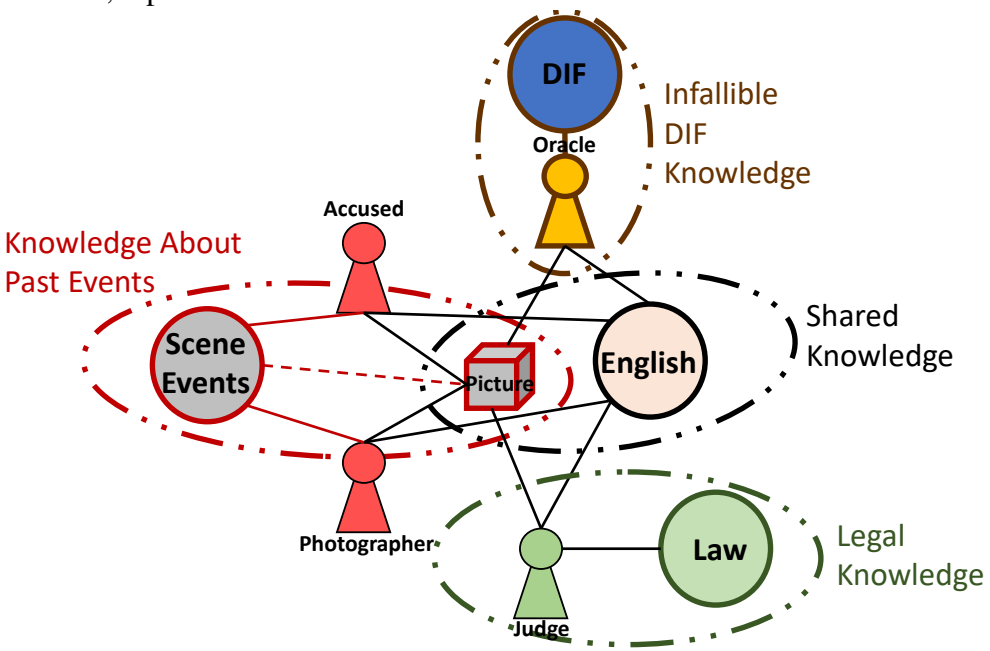


Figure 5.3: Representation of knowledge distribution and access to information in a hypothetical court. In the center, both the picture and knowledge of English represent shared knowledge among all parties. The knowledge of past events, in red, is limited to individuals present to such events. On the top, in brown, the infallible knowledge of DIF is accessible through the oracle. The judge is the only assumed to have legal knowledge in this case, represented at the bottom.



- The photographer: the person who took the picture in question. (S)He has knowledge of English, has access to the picture, and has past knowledge about the scene

events. This knowledge is subject to his limited perception of the scene and interpretation;

- The accused: the person described as being in the scene and disagreeing with the photographer's account. He is assumed to know English, and has at least at one point had access to the picture, as he denies what is represented in it. Since he was a participant in the events, he also has knowledge about them, limited by his perception and interpretation.

The conflict arises from the fact that both parties that share a common knowledge of the events disagree. The mediating party, the judge, cannot trust any of them because there are conflicts of interests. Using his legal and criminalistics knowledge, however, he might solve the conflict by understanding motives, using jurisprudence, and analyzing evidences. In a real court, one would also expect the presence of lawyers, attorneys, prosecutors, and even a jury, which are considered important parts of due process. They can provide additional knowledge and viewpoints, but they also make the process more complex.

The actors in this example have been chosen to provide flavor and context, while preserving simplicity. The core of the issue is a conflict between two parties, mediated by a judge. In the process of analyzing the key piece of evidence, the picture, the judge brought in the infallible DIF oracle. The only shared information between all the parties are the picture itself and a limited knowledge of English.

Since the oracle was not able to solve the Judge's problem, such problem must lie outside of the oracle's sphere of infallibility. Therefore, it must lie either in the knowledge of the scene events, law/criminology, or English. In the following subsections, I explore the possibility of providing these additional spheres of knowledge to the oracle, and how it affects the problem at hand. I show that if the infallibility of the oracle could extend to any of these three spheres, it could solve the problem. The implications of doing so, however, are absurd or implausible. This exercise is fundamental to our discussion on both the epistemology of DIF, and its objectives as a research field.

5.4.1 Knowledge of Past Events

The first possibility to be explored is knowledge about past events (red area in 5.3), moving "Scene Events" within the oracle's infallibility area (brown area). If the DIF

oracle had complete knowledge about the scene events, he would be able to tell if the picture corresponds to reality or not. Therefore, the judge would be able to accept the picture as evidence, and rule accordingly. This is not a valid solution, for more than one reason:

1. It assumes that knowledge about a scene in the past would qualify as DIF knowledge, which is our requirement for infallibility under the oracle;

2. If knowledge of the past could qualify for infallibility, there would be no need for DIF. The judge could simply ask questions about the past;

3. Both the photographer and the accused have accounts of the past limited by their senses and interpretation. Infallible knowledge implies surpassing these limitations.

To assume that the DIF oracle could have infallible knowledge of the scene would virtually put it in an omniscience position. Therefore, it is not reasonable to consider knowledge about the past to fall within the sphere of DIF infallibility (Fig. 5.3a). We know for sure, however, that a great deal of scientific knowledge is based upon knowledge of past events. All scientific experiments are events, and building scientific theories requires using information from experiment that happened in the past (Fig. 5.3b). Therefore, all scientific knowledge is based on knowledge of past events. When defining our borders for DIF infallibility, how should one proceed?

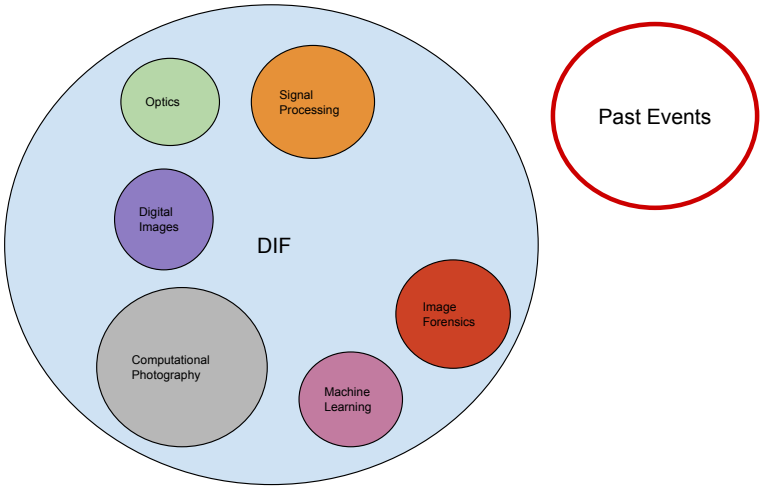
This is, in fact, a problem not only of DIF, but of scientific realism and philosophy of science. If the DIF oracle has infallible knowledge about the behavior of some natural phenomena, it could answer questions about the nature of reality itself. Instead of designing experiments, the job of scientists would be to design questions to the DIF oracle.

If it is not possible to estimate bounds to the knowledge of DIF from a perspective of time and empirical observations, then it is not conceivable to consider bounds for its infallibility. This is a problem of epistemology of modality. However, having a constructive empiricist view of DIF provides a solution.

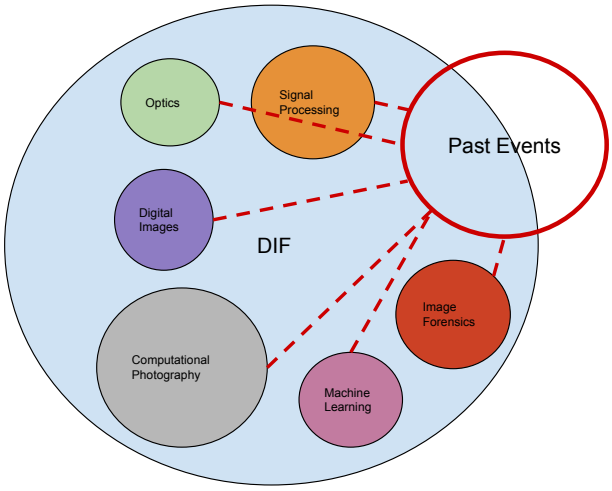
Constructive empiricism (FRAASSEN, 1980) states that science does not hold truth over reality, but rather empirical adequacy. In this sense, an infallible DIF oracle can have some infallible knowledge about past events, but only in the sense that they form a basis for the empirical adequacy of DIF. One way to think about this is that the oracle has access to infinite plausible scientific papers in DIF. When performing an experiment, a

Figure 5.4: Dependence of DIF knowledge on past events.

(a)



(b)



scientist never measures all there is to measure, and much of the information never reaches the final paper. The final product is a concise synthesis of experiments and theory, and many successive papers build an empirically adequate explanation of the topic. However, from the papers alone it is impossible to reconstruct the event of the experiment.

5.4.2 Law and Criminology

As discussed in the Introduction, one of the definitions of DIF was inside the forensics sciences (Fig. 1.2), which is closely related to law and criminology. Why is it not reasonable to assume, then, that the oracle has access to infallible knowledge about these subjects? Let us recall the following interaction between the oracle and the judge:

- J: Can image 1 be considered a valid piece of evidence in this court?

- O: I do not know what qualifies an image to be considered a valid piece of evidence in this court, your honor.

What qualifies as valid piece of evidence in a court depends on knowledge of the ruling law, which depends on place and time. Ultimately, it could be a decision ruled by the judge. If the judge had asked "Can this image be considered a valid piece of evidence in **some** court?", an infallible answer would be yes. If the question had been "Can this image be considered a valid piece of evidence in a German court?" To be able to answer such a question, the oracle would need knowledge of both German legislation and rulings. Even if we consider that the oracle has memory of all cases in Germany involving DIF, no one of the cases were identical to this one. They involved different people, different pictures, and different premises. Being able to extrapolate from similar cases requires an underlying understanding of legal process, which arguably falls outside the scope of DIF.

If we try to change the structure of the problem by expanding DIF to law and criminalistics, we must accept there is an oracle that is able to perform infallible judgment, and therefore have no further need for judges. Let us assume the opposite, however. Temporarily, and only for this case, the judge has access to infallible DIF, and we can read the transcript of his/her thoughts:

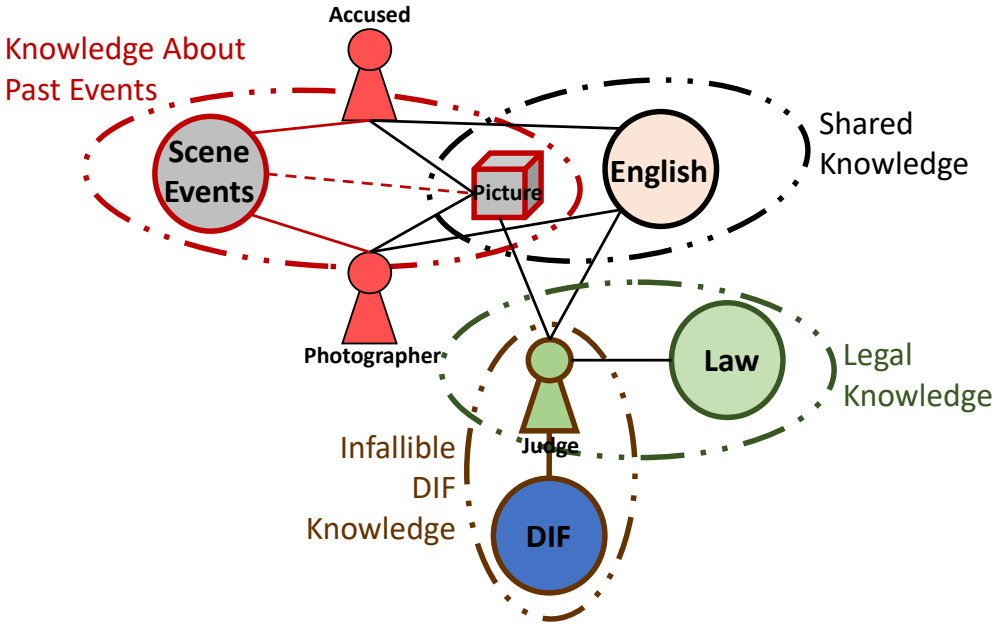
- J: Can image 1 be considered a valid piece of evidence in this court?

- J: I suspected that the colors were strange, but according to the shadows this picture could have been taken during the sunset. There seems to be no signs of integrity violation on the file itself, since the coding and compression traces are compatible with the reported by the camera manufacturers.

- J: Seeing no signs of wrongdoing or ill-will, I must accept image 1 as valid evidence.

Removing the DIF oracle from the court and directly endowing the judge with DIF infallibility would result in the structure of knowledge represented in Fig. 5.5. Notice how

Figure 5.5: Restructuring of knowledge distribution described in Fig. 5.5. The oracle was removed, and his infallible knowledge in DIF was transferred to the judge.



the previous limited knowledge of the judge in Law and Criminology remained the same, but he was able to solve the problem. In some sense, he was able to solve the problem because it was **his** problem. It is not required by a judge in a court of law to be infallible, therefore an infallible answer was not required. He is accountable by his duty to society, which is covered by his legal knowledge, and by his fairness, which is covered by his personal judgment.

With access to infallible DIF, the judge could have dissected the image for any minor traces of forgery. He could have made statistical analysis on the probabilities of the colors of the hair of the subject to really be caused by sunset light, and so on. This is not only unnecessary, but unwanted from his point of view. We have already established that we do not possess infallible knowledge from the past, because some information is always lost. Gathering more data can further complicate the decision process. The judge has studied the case, knows the involved parties and similar cases, and balances this in his decision and thoroughness of analysis.

If we recall the difference between Worlds 1.0 and 1.1, while both had access to infallible DIF knowledge, only the latter developed an “oracle” capable of holding all knowledge. It seems intuitive that World 1.1 has a clear advantage, but this example shows otherwise. With this approach the problem was solved without adding additional knowledge to the system, simply removing the oracle from between the infallible DIF and the judge. Furthermore, the judge used a very limited subset of the vast possible

knowledge to reach a decision.

An external observer analyzing similar court cases from Worlds 1.0 and 1.1 could find that, without the oracle, things actually run smoother. By not employing artificial intelligence (AI), agents that needed DIF would study from the infallible source and gather limited but applied knowledge. World 1.1, by delegating all the responsibility of holding knowledge to the AI lacked practical agents.

This conclusion seems odd, as there is nothing limiting agents from World 1.1 to study and learn themselves, achieving at least the same amount of success as World 1.0. What is then, the point, of having an oracle? One could argue that with the oracle it would be easier to train agents and access the knowledge, giving an advantage to World 1.1 over World 1.0. If both worlds adopt training procedures and good learning practices, having an oracle is better than not having. This is a valid point, but the utility of such an oracle lies within one of the fundamental problems of DIF: language.

5.5 Language

Since it is not sensible to consider English to fall within the sphere of infallibility of DIF, it was left out. This would also be the case whatever the language in question is, because if English cannot be considered part of DIF, neither should Portuguese, German, or any other language. This creates another interesting paradox about the limits of infallibility. A constructive empiricism approach has been adopted to solve our issues with being infallible about the past, and the used analogy was that the oracle had access to infinite scientific papers on DIF. All those papers should, invariably, be written in some language. To be able to use this knowledge in practice and answer questions about it infallibly, it holds that the oracle should have infallible knowledge of that language.

One way to solve this paradox is to assume that all knowledge inside the oracle is stored in a different, infallible language, such as a formal meta-language. When someone asks a question in English, it is translated to the meta-language, processed infallibly, and then translated back to English. This is very fitting with what can be read in the transcript, as the oracle asked to the judge define what a “real image” is. The Judge, complying with his linguistic specificity, tried to explicitly define what a real image was to him. The result was an answer not devoid of meaning or information, as it described many fundamental aspects of the subject in question, but nevertheless not useful for the problem.

The Austrian philosopher Wittgenstein once said that “whereof one cannot speak

thereof one must be silent", and his work on language is very fitting to this discussion. Since the oracle does not have infallible knowledge of English, as this is not DIF per se, it cannot guarantee that the infallibility of its answers will hold after a translation process. Instead of being silent, however, it provides tautological answers that carry information, but not the requested information.

What would be the implications of removing this language barrier, then? There are two evident options: giving the judge knowledge of the DIF infallible meta-language, or giving the oracle knowledge about English. The first option, surprisingly, is already described by Fig. 5.5. Without entering in discussions about AI consciousness, giving the judge the ability to articulate his thoughts in the infallible DIF meta-language, then querying the oracle for answers is practically equivalent to endowing him with infallible DIF knowledge.

The prospect of teaching humans an infallible DIF meta-language also raises an interesting question about the differences between Worlds 1.0 and 1.1. It seems that bypassing the language barrier, such as English, is an advancement for World 1.1, therefore desirable. If both worlds have similar storage structures for the DIF information, such as physical servers, the only difference between both worlds is the supposed consciousness of the oracle. If the meta-language is only capable of describing statements pertinent to DIF, an agent wishing to ask questions to this "consciousness" will filter any extra information. There seems to be no difference, then, between a programmer in World 1.0 querying a database where the infallible DIF knowledge is stored, or a person in World 1.1 constructing a question in the DIF meta-language for the oracle.

A parallel can be drawn between the different levels of representation developed in our planning solution (Fig. 4.6). The high-level domain language used to describe the forensics knowledge is not how the planner "thinks", neither is our intermediate level of representation: its language is PDDL, constrained by a formal, logical structure. It is through our intermediate representation that we are allowed to overcome the limitations of PDDL, and it is through our high-level representation that users are allowed to express their knowledge in the solution.

The second option for solving the language problem would pull knowledge of English inside the sphere of infallibility, which is equally unsatisfying. Again, we fall into the pitfall of defining boundaries of knowledge. What does having infallible knowledge of English means? Does it entail all forms of English, such as American English, Scottish, etc? Even if we stretch to fit the whole Indo-European language family, there is still

the need to account for dialects, slangs and context. To have knowledge about all the possible meanings of an expression and still be infallible in its evaluation requires complete knowledge of context.

The transcript of the conversation between the DIF oracle and the judge would have been the same if the oracle had infallible knowledge of the English language. Even within "our world's" DIF literature there is no clear definition of what could be a real or a fake image. The judge would have to be as specific as possible on his meaning and intentions, just as in the example. If one is to recall the transcript of the judge's thoughts when (s)he had access to infallible DIF knowledge (Fig. 5.5), they might seem simple to communicate, but they are based on his knowledge of the case, law, and criminalistics, too. For the DIF oracle to be able to replicate her/his train of thought, it would essentially mean to read the judge's mind, even with an assumed infallible knowledge of the English language.

5.6 The Problem with Probability

So far a great deal of our criticism of DIF and the oracle hinges on truth statements and our notions of infallibility. We are expecting the oracle to give an absolute, binary answer, to questions that either require inaccessible knowledge or are non-binary. What if, instead of providing absolute, infallible, truth statements, the oracle provided probabilities? Our current literature on DIF already expresses a great deal of results and knowledge using probabilities and statistical distributions. Furthermore, the judge is the final authority, and his decision might as well be based on infallible probabilities than on infallible truth statements.

Even if the Oracle is able to provide probabilities, it is still unable to reach the past, or define the intentions of the judge when asking if an image is "real". Additionally, now there is the issue of the different meanings for probability. There are many interpretations for probability, and they are commonly used interchangeably or without proper specification (HáJEK, 2003), causing confusion.

To illustrate these issues, we will recall the *Sleeping Beauty problem* (ELGA, 2000) from decision making, showing how a simple question about the probability of a coin toss has many valid answers. We show that without knowing prior probabilities of events, probability estimations tend to be vague by prioritizing general cases. We then construct an example of probability estimation for forgery in images, showing how this

issue can lead to bias.

5.6.1 The Sleeping Beauty Problem in DIF

In this problem, Sleeping Beauty (SB) agrees to participate of an experiment. The experiment takes place in four days, from Sunday to Wednesday. On Sunday she will be put to sleep, and a coin will be flipped. Regardless on the result of the coin, she will be waken up on Wednesday and the experiment is over. The coin is fair, meaning it has 50% of chance of coming out either heads or tails, and it will be tossed on Sunday after she is put to sleep. The result of the coin determines how the experiment will proceed in the following days. If the coin turns heads, she will be waken up on Monday, interviewed, and put back to sleep with her memory erased, until she wakes up again on Wednesday and the experiment is over. If the coin turns tails, the same thing will happen on Monday, but instead of spending Tuesday sleeping, she will be waken up, interviewed once again, have her memory erased, and put to sleep until Wednesday. During the interview, she does not know which day of the week it is, and is asked: "what is your credence that the coin has come up heads?"

Figure 5.6 describes the structure of the experiment. 'H' and 'T' stands respectively for heads and tails on the coin, which is tossed on Sunday after SB is put to sleep. 'A' stands for awoken and interviewed, which will happen on Monday regardless of the result of the toss, and Tuesday only if the coin landed tails. On a heads toss, the 'S' on Tuesday indicates she will be left sleeping. Essentially, the difference between heads and tails is the number of times (once or twice) that SB will be waken up, interviewed, and have its memory erased. In any case, she is asked the same question and has no ability to tell which day of the week it is.

Since the coin is fair and SB knows this, it seems obvious that her answer should be 50%, or $\frac{1}{2}$. There are only two possibilities, heads or tails, and each one is equally probable. This is shown on Figure 5.7 (left), along with the probabilities for being in each awakening instance. If there is 50% chance for tails, and probabilities add up to 1, then when waking up there is a $\frac{1}{4}$ chance of being in either tails awakening. This is one of the "solutions" to the problem, but it ignores the experiment itself! By knowing the coin is fair, this answer could be given on Sunday, before Sleeping Beauty is put to sleep, and there is no need to even knowing the experimental setup.

Another solution focus on the agent herself, as she tries to answer the question

Figure 5.6: Description of the Sleeping Beauty experiment according to the two possible outcomes of a coin toss. 'A' stands from awoken, 'S' for sleeping, and 'E' indicates the end of the experiment.

	Mon	Tue	Wed
H	A	S	E
T	A	A	E

Figure 5.7: Description of the Sleeping Beauty experiment according to the outcomes of a coin toss. The probabilities outlined in orange in (a) and (b) correspond to the "halfer" and "thirder" interpretations of the problem, respectively. In (a), the probability distribution of the coin is taken into account ($\frac{1}{2}$ for both heads and tails). In (b), the probability of each individual awakening is taken into account, arriving at $\frac{1}{3}$.

(a)

	Mon	Tue	Wed
H 1/2	A 1/2	S	E
T 1/2	A 1/4	A 1/4	E

(b)

	Mon	Tue	Wed
H	A 1/3	S	E
T	A 1/3	A 1/3	E

with limited information when awaken. There are in fact 3 possible awakening events, of which only one happens on a heads result (Figure 5.7, right). She does not know which day it is, if she has already been woken up, or if she will be once more. Therefore, she will always provide the same answer, and this answer should maximize the probability of being right. In this case, it is logical to answer that her credence the coin has come up heads is $\frac{1}{3}$! If the question was "has the coin come up heads?" and she answered "yes", she would be right on $\frac{1}{3}$ of the awakening cases. The first solution is called the "halfer" position, and the second is called "thirder", but more positions arise with variants of the problem.

One of the differences between the "halfer" and "thirder" positions is that the latter takes in consideration the setup of the experiment, which is a kind of information. If the

setup was different, and on a tails result implies that SB is woken twice on tuesday instead of once, the "thirder" position would say now that there is $\frac{1}{4}$ credence the coin came up heads. This would happen because there are now 4 awakening events instead of 3, and 3 of them happen on a tails result (one one Monday and 2 on Tuesday). With only 1 possible awakening in the heads case, the "thirder" response would change from $\frac{1}{3}$ to $\frac{1}{4}$ in this new experiment, while the "halfer" answer would remain $\frac{1}{2}$. If the experimental setup consisted of 999 awakenings on Tuesday for tails, instead of 1 or 2, then the "thirder" answer would be the shocking $\frac{1}{1000}$ credence for the coin to have come up heads, even if we are still considering a fair 50/50 coin. In this case, SB would be ignoring the a priori toss chance in favor of being right most of the awakenings. This is similar to a classifier that classifies everything as "class 1": if the frequency of occurrence of "class 2" is $\frac{1}{1000}$ of the frequency of "class 1", it will be right 99.9% of the time, having an almost perfect accuracy, but very little usefulness.

The knowledge of the structure of the experiment is a sort of *static* information, as it is presented to SB on Sunday, but it was demonstrated that can affect her answer. Consider now the case that when SB is waken on Monday, she is told that it is Monday. Regardless of the coin toss she will be woken on Monday (Fig. 5.6), so it seems that the information that it is Monday should have no effect on her answer. However, the key information is that it is not Tuesday, making the case similar to the Monty Hall Problem (SELVIN, 1975), and the optimal answer to be $\frac{2}{3}$. To understand why we can think about the question as "if you answered that the coin turned up as heads, what is the chance you are wrong?", and then subtract that chance from the full probability. The only chance for SB to be wrong is if it is Monday and the coin landed tails. Therefore, to maximize her chances of being right using this information she must change her answer to $\frac{2}{3}$ (*i.e.*, $1 - \frac{1}{3}$). This is another counter-intuitive result that showcases how credences tend to distort the original question in favor of an answer that is right in the majority of scenarios.

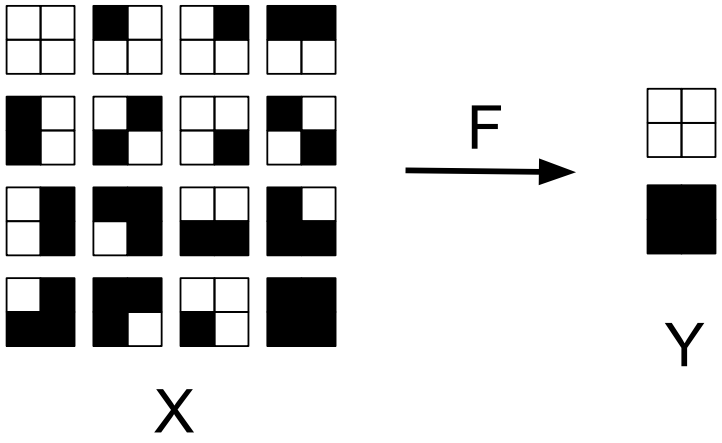
This problem exposes a conflict between different interpretations of probability, or between absolute and subjective probability (also referred to as belief or credence). We can imagine the DIF oracle as Sleeping Beauty being awakened for a trial and interviewed. An image is shown and the oracle is asked: "what is your credence that this image has been forged?". The image was forged in the past, before it was awaken, just as the coin toss takes place on Sunday, before the first possible awakening for SB. A probable infallible answer for our oracle would be $\frac{1}{2}$, since the image either has been forged or

not, but this is not useful at all for the judge! A “thirder” answer to this case would entail estimating the combined probability of all possible ways that image could be forged.

5.6.2 Estimating Forgery Probabilities

To understand what it would mean calculating a priori probabilities of an image be forged, we can construct a simple example. Imagine a 2×2 image, where the pixels could be either black or white. There are 2^4, or 16 possible images at this resolution representing the domain X. These images can undergo filtering operation F, which results in 2x2 images of the same color, either 4 black pixels or 4 white pixels, representing a codomain Y. Figure 5.8 illustrates both the domain X, the filtering operation F, and the resulting codomain Y of output images.

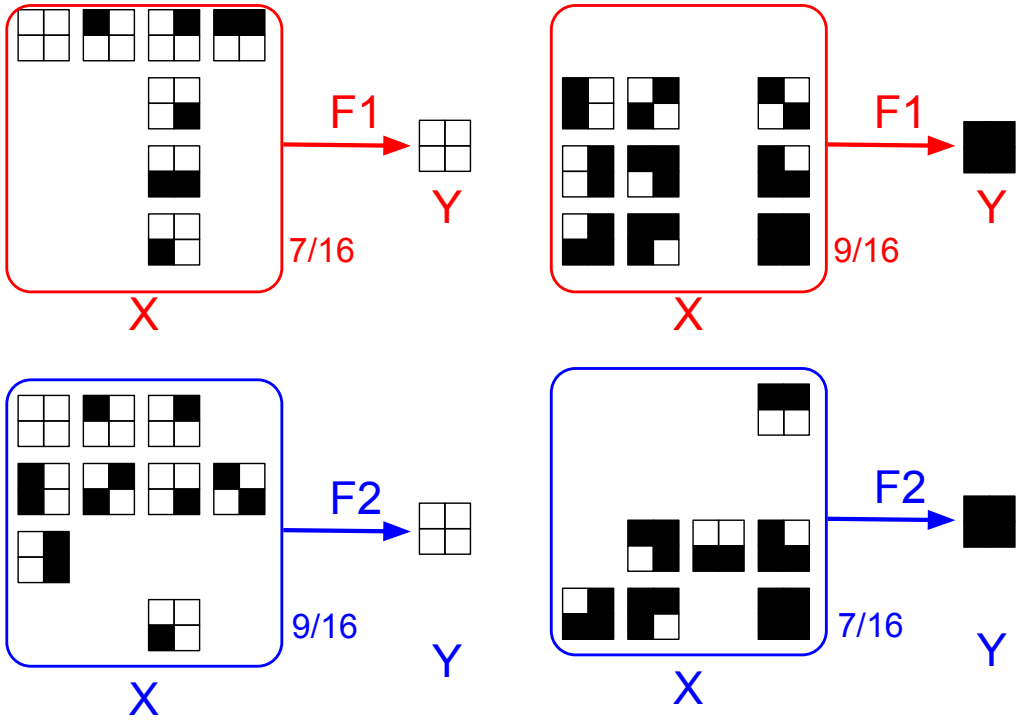
Figure 5.8: Example of a domain X with all possible 2×2 binary images, and a transformation F that outputs either a white or black 2×2 image.



Let us describe each pixel as P_{tl} for top-left, P_{tr} for top-right, P_{bl} for bottom-left, and P_{br} for bottom-right. There are two possible filters, which are combinations of and ($\wedge$) and or ($\vee$) operations between the 4 pixels. Filter 1 is $(P_{tl} \wedge Ptr) \vee (P_{bl} \wedge P_{br})$, which can be translated as “if either the two top pixels or two bottom pixels are white, the resulting image will be all white, black otherwise”. Filter 2 is $(P_{tl} \vee Ptr) \wedge (P_{bl} \vee P_{br})$, which can be translated as “if both the bottom part and top part of the image contains at least one white pixel, the resulting image will be all white, black otherwise”. Figure 5.9 shows the mapping between X and Y for both filters, along with the amount of combinations that could produce each Y, which could be considered the priors. We could say with $\frac{7}{16}$

credence that any random 2×2 black and white image will become completely white after going through the F1 filter.

Figure 5.9: Domain and co-domain for the example transformations F1 and F2.



Let us assume Y are the images that could possibly be presented at the court, and X are all the possible scenes that could have originated them. We could now define different specifications for what it means for a 2×2 image to be forged, for instance having been originated from a "checker" pattern. The prior probability of any image being forged is $\frac{2}{16}$ now, because there are 2 possible "checker" patterns, but this does not help the judge. The judge wants to know the credence for this image to be forged. This deviates from our original SB problem proposition, as we are giving the oracle additional information during the interview (an Y image to be analyzed). Additional information should not affect the a priori probabilities of an event, or how it happened, but it does affect credence.

The knowledge of it being Monday or not in the SB problem is different from the knowledge of the structure of the experiment, as it is obtained only on Monday. This type of knowledge is similar to an image being presented to the DIF oracle at-time in court. Instead of "Given that it is Monday, what is your credence that the coin turned up heads?", one asks "Given this image, what is your credence that it has been forged?". One cannot determine if the image has been forged using only information present in

the image itself, unless there is some nonsensical element that rules out the possibility of it being not-forged (for instance, a person with horns and green skin). In the same way, the information that is Monday is not enough to determine the coin toss, but if the given knowledge was that it was Tuesday, SB can have 100% credence the coin has come tails. This is a consequence of the problem setup: it is a necessary condition for her to be awaken on a Tuesday that the coin has come up tails.

When the oracle is given a 2×2 white image (Y) to analyze and considering a previously "checkered" image to be forged, there is still the question of which filter was applied. Observing Fig. 5.9 it can be noticed that a "checkered" pattern will only generate a white image when using F2. The question then becomes, "what is your credence F2 occurred, given that Y is white?", which is a very dangerous line of reasoning for a court of law! The Y image being white and the F2 transformation are not dependent events, and neither one individually implies forgery or wrongdoing in the image. Our 2×2 X domain will generate a white image with F2 $\frac{9}{16}$ of the time, $\frac{2}{16}$ more than with F1 ($\frac{7}{16}$). This would imply that, by definition, we should be more suspicious of white images being forged, even if they represent $\frac{1}{2}$ of the Y domain. In the described universe of 2x2 images with transformations F1 and F2, white images will always have a higher credence of being forged rather than not forged.

On Section 4.3.1 we exposed a set-theoretic formalization for DIF, where digital images can be mapped to a countably infinite set $I^Z = I^M$, or countably finite for a determinate resolution, bit depth, etc. This example (Figs. 5.8 and 5.9) is developed from our experiments with such formalization, and represent a possible scenario over $I_{2,2,2,1}$. It shows that being able to estimate all possible a priori possibilities for image transformations is not necessarily useful.

The problem with any credence value the oracle gives is that it avoids the actual issue of the forgery, maximizing his odds of being right based on the structure of the problem, the same way SB ignores the actual toss of the coin. Since she is being awaken on Monday for both cases, the coin could be very well be tossed on Tuesday morning to decide if an awakening will happen. If she does not know the coin will be tossed on Tuesday morning, her behavior will not change but she will be giving answers about a future event on Monday, which is strange. If she does know about this, but is not told when it is Monday, her behavior will not change. If she knows about this, and is told it is Monday on Monday, her answer would change from $\frac{2}{3}$ to $\frac{1}{2}$, since the question then becomes "what is the credence that a fair coin that will be tossed in tomorrow morning is

heads?”.

For the judge to be able to use the credence values given by the DIF oracle, he would need to understand the whole structure of the problem, and the oracle’s answering strategy. Is the oracle a “halfer”? If so, it is by definition ignoring the image in question and giving priors. Is it a “thirder”? If so, its answer is only as good as the capacity of the judge to interpret its reasoning, which again falls into a knowledge sharing and communication problem. As we add more resolution, channels, color depth, operations, and the ability to perform chains of operations, it becomes almost impossible to keep track of all possibilities. This is the case with modern images, with the added difficulty of not being able to define what a “forged” or “real” image is.

5.7 Summary

The ideas exposed in this Chapter might seem forcibly pessimistic, undermining any possibility for satisfactory solution for DIF problems, but this is not the case. To show the limits of technology, we have purposely constructed scenarios which avoided addressing the issues of both communication, and access to knowledge. What restricted the judge to determine if the image could be used as evidence was his inability to properly communicate with the oracle. His necessity for communicating with the oracle, in turn, came from his inability to access knowledge: either about the truthfulness of the witness testimony, or of the actual events.

In the Introduction (Section 1.2), I discussed different interpretations and scopes for the field of DIF, based on the state-of-the-art. Two of those interpretations were based on hierarchical descriptions of related research fields (Figs. 1.2 and 1.3). The infallible knowledge of the oracle was based on such perspectives (Fig. 5.2), as it allowed us to play with the concept of "in", and "out" of scope. Everything that we consider to be inside DIF could be endowed with the miracle of infallibility, but, in the end, the judge’s questions were out of scope. Had the judge asked technical questions such as "What was the jpeg quantization table used in this image?", things would have been different.

If we consider DIF purely as a technical field, we are focusing on the "Digital" aspect, which is only part of the problem. Most practical questions that one would want to ask to DIF would automatically fall outside of this scope. People care ultimately about the image content, which means they care about the communication of the content. The "Digital Image" happens to be the means by which the content is being stored, and the

role of "Forensics" in this case is to provide trust protocols. This is the view of DIF that I advocate in the Introduction (Fig. 1.4), where each component word "Digital", "Image", and "Forensics" generalizes rather than specify the scope. The cases of conflict involving the oracle would not have happened had we used this definition of scope. It would also not be a fair comparison, because this definition implies an almost unbounded range for the infallibility of the oracle.

The likelihood of a technological oracle existing in the future is not important for the discussion. What should be taken from this is that technical DIF knowledge is limited by how well we can understand it and use it to mediate communication. There is a massive amount of technical research on DIF (Chapter 3, a great deal of which society has not been able to understand and articulate yet. One of the reasons for that is the difficulty in forming coherent epistemological foundations for DIF, which is discussed in the next Chapter.

6 KNOWLEDGE UNDER DIGITAL IMAGE FORENSICS

This Chapter analyzes the epistemology of Digital Image Forensics, both as a research field and as a framework for investigation. The two main subjects of discussion are the nature of evidence, and the models used to describe and reference elements of knowledge withing DIF, such as properties. These two subjects are generally intertwined in philosophy. In the case of DIF, they are specially entangled due the technical and investigative aspects of the field. Furthermore, we argue that it is hard to separate the empirical evidence we use to build our knowledge and techniques, from the objects of our forensics analysis.

One of the most classical definitions of knowledge is the JTB formulation, which stands for justified true belief (KAPLAN, 1985). It poses that an agent S **knows** a proposition P iff:

1. P is **true**,
2. S believes that P is true, and
3. S is justified in believing that P is true.

The first requirement for something to be considered knowledge is that it be true, so the first condition talks about the real-world state of proposition P. The second condition is that S believes that P is true, which is not a statement about the real-world, but rather about the internal state of S: if S does not believe in P, it cannot constitute knowledge. With these two conditions, the real-world value of P and the mental state of S match, but this is not enough to constitute knowledge. The third condition, which is the justification clause, states that S only **knows** about P if (s)he is justified in doing so. The nature of this justification, however, has been an ongoing debate in the field of philosophy for a long time. To understand why, let us instantiate an example:

1. P = It is 2:30,
2. S believes that it is 2:30, and
3. the reason S believes that P is true is because S looked at the the wall clock and it is marking 2:30.

It seems fair to say S knows that it is 2:30, and we are inclined to accept this justification on the grounds that it is actually 2:30. However, a careful analysis of the

third clause shows that this is not enough to ground the belief. It just so happens the clock S is broken, and has stopped at 2:30 in the past. By coincidence, however, it is 2:30. It seems wrong to consider this as knowledge, because had S looked five minutes before (s)he would believe it was 2:30 when it was actually 2:25. This sort of counter example that challenges justification for beliefs is called a *Gettier case*, famously presented by Edmund Gettier in his short paper "Is Justified True Belief Knowledge?" (GETTIER, 1963). Gettier cases are important for any discussion on DIF, as they challenge our notions of what constitutes evidence and justification.

P can be a composite proposition where inference applies, too. In the original Gettier case there are two people, let us call them Johnny and Mark, who are on a waiting room after a job interview. Johnny is bored and starts counting the coins in his pocket, showing Mark he has ten coins. At one point, the president of the company joins the room and compliments Mark on his great interview. From Johnny's perspective, this is evidence that Mark will get the job, and he thinks to himself "the man with ten coins in his pocket will get the job". This is inferred from $P =$"Mark has ten coins in his pocket", and $Q =$"The man with ten coins in his pocket will get the job". Since Johnny has evidence for both P, and Q, it holds. However, it turns out Johnny is hired for the position and, unknowingly to him, he also had ten coins in his pocket. Therefore, the man with ten coins in his pocket really got the job, but it is wrong to say that Johnny **knew** about it.

6.0.1 Evidence-based Justification

In the user study reported in Chapter 2, we required subjects to provide evidence (by pointing to it) whenever they thought an image was fake. For our experiment, that accounted for the third clause: justification. If an image is fake and the subject answers it is fake, we have a true belief (clauses 1, and 2), but not a justified true belief. We introduced the concept of evidence masks (Fig. 2.2) to determine if the location provided by the subject could be considered valid evidence. In the trophy example, we accepted points around the spliced trophy and around the missing shadow as evidence (Fig. 6.1).

Figure 6.1 shows all the provided evidence points for this example, with the edited area highlighted. The majority of subjects provided valid evidence, but in different ways. These positions represent reasoning, the justification for the belief that the image had been edited. Each of those points can be translated into a different statement:

Figure 6.1: Heatmap of points provided as evidence on the trophy image (Fig. 2.2). The white outline on the trophy marks the editing mask.

- I clicked on the trophy because it is the edited object;
- I clicked on this specific part of the trophy because its lighting is odd;
- I clicked on the wall because there should be a shadow cast here;

- I clicked on the border of this object because the color transition seemed wrong.

During the experiment's pilot test, we observed a variety of strategies subjects used for justification, and for this reason crafting evidence masks was not a simple task. We had to carefully consider different types of reasoning that could visually expose the editing in each image, and how "mouse clicks" could be used to express it. Some people chose to point to the pedestal, instead of the missing shadow or the trophy, but it is impossible to know if they really missed the forgery or just were mislead by false evidence. An extreme case of misleading false evidence can be seen on Fig. 6.2, where the right slipper has been spliced in the image. The majority of subjects, which had a true belief that the image had been edited, however, chose to point at the left slipper as evidence, and were considered wrong in our metric. A probable reason for the subjects' mistakes is the visual confusion caused by the lighting. At a first glance, it seems the lighting on the right slipper is more consistent with the reflection on the bottom-right, and the left slipper contrasts with it.

Figure 6.2: Heatmap of clicks provided as evidence on a splicing image in our subject test. The right slipper was added to the image, but the left one received most of the clicks. The white outline on the right slipper marks the editing mask.

If we were to transfer this scenario to a practical case, such as a court of justice, there could be serious consequences. A ruling based on dubious evidence can dismantle a

case, and open precedents to dispute similar cases. This can be specially dangerous when there are adversarial sides confronting, as the unjustified true belief can become the focus of the opposite side's strategy. It is not easy, however, to decide what accounts for proper justification. In fact, since Gettier put in question the JTB formulation many researchers have come up with different solutions to patch this hole in our definition of knowledge. Some of these approaches are discussed further in the text, but first we provide more ways in which DIF is prone to Gettier cases.

Gettier cases happen when unobserved phenomena voids our justification, or when they are the actual cause of our true belief, rather than our supposed justification. On the clock example, the fact that the clock was broken invalidates it as a time-measuring device. On the job interview example, the fact that Johnny had ten coins in his pocket and his superior performance to Mark in the interview can be seen as causes for "the men with ten coins in his pocket will get the job" to be true. The fact that Johnny had ten coins in his pocket, as well as the company's decision process for hiring were both unobserved phenomena.

In DIF, however, it is hard to account for all unobserved phenomena. For one, because an image only contains limited information, and it is impossible to reconstruct its past (Section 5.4.1). Secondly, because the statistical mappings of image features often result in numerical statements that are hard to scrutinize. This was slightly breached in Section 4.3.2, where we discussed dividing the space of possible images according to their properties. Most forensics techniques use some form of supervised learning to perform classification by partitioning some high-dimensional space. Thus, the way supervised learning techniques subdivide such high-dimensional spaces is based on examples, rather than on an implicit or constructive definition of a property. For this, one builds a dataset containing both positive (images that do have a certain property) and negative (images that do not have the property) cases. The way in which we **know** the positive cases happen to contain the property is because we specifically constructed them, or that it was guaranteed by some other external source or measure.

6.1 Properties are too abstract, Traces are too concrete

If *jpeg* is a property some images can have, how is it defined? Does an image acquire this property after undergoing *jpeg* compression? If this is the definition, foolproof **justification** for believing an image has this property requires knowledge about its history

as an artifact, and can never be guaranteed only based on its content as a numerical entity on the domain I. This is an unpractical definition for most ends, so it is avoided. The meta-data such as the EXIF cannot be trusted, since it can be easily changed and ill-will cannot be ruled out. Rather, what is usually done is a statistical modeling of the numerical content of images that have undergone compression. The property is re-defined in terms of the closeness in a hyper-dimensional space to positive samples.

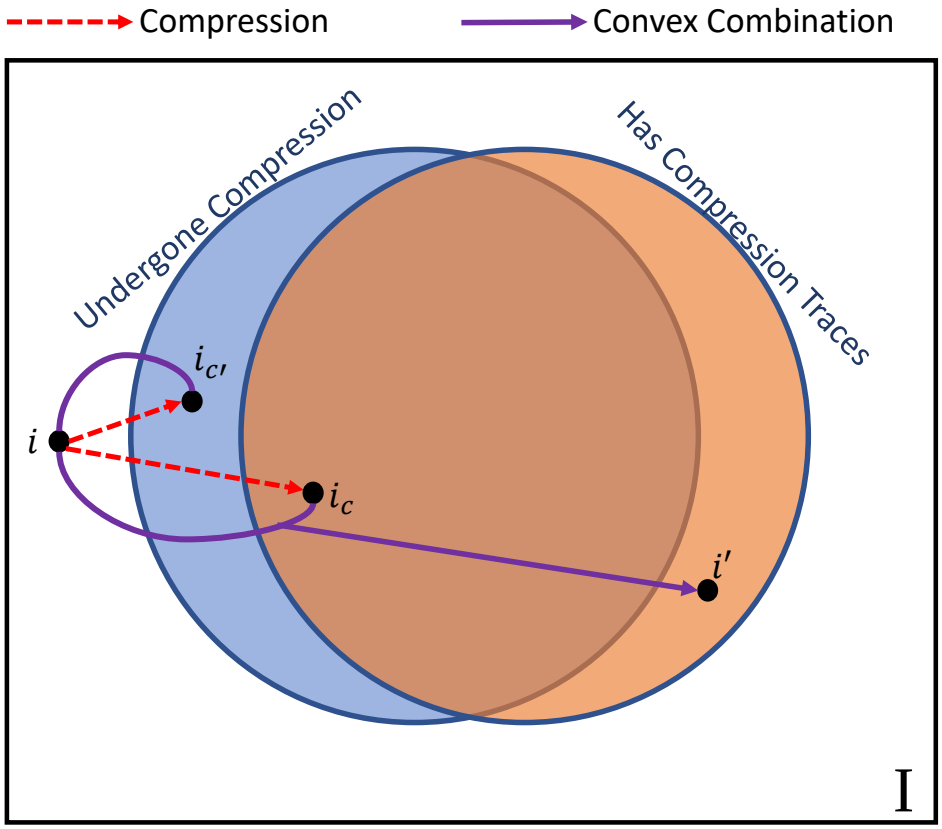
We use the term *property* loosely here on purpose, because there is no formal definition on the literature. Indeed, much of the confusion is avoided by saying, for instance, an image has jpeg compression traces (Section 3.2). The point of whether an image is *truly jpeg* becomes rather platonic in favor of a more technically practical interpretation. The issue, then, is how these technically practical interpretations can be used as basis for justification of knowledge. Our quantitative experiments with transferring JPEG compression traces (Section 3.4.2), for instance, show that it is possible to inject them into specific parts of the image without changing the content. From an uncompressed image i one can obtain a compressed version i_c, and there is an α such that $i' = \alpha i + (1-\alpha) i_c$ will show traces of compression, as evidenced by our experiments (Figs. 3.4a, and 3.4b). This is a simple form of feature injection (IULIANI et al., 2014) that illustrates the complexity of translating something as *having undergone compression*, or even *having traces of jpeg compression* as numerical statements.

6.1.1 A Dualism in Reference

Using this formulation it is possible to construct images in the form $i' = \alpha i + (1-\alpha) i_c$ that will test positive for traces of compression, even without having undergone compression. Figure 6.3 illustrates this, and shows that, with such a procedure, it is possible to achieve any of the four possibilities: (i) images with no compression that show no compression traces; (ii) images with compression that show no compression traces; (iii) images with no compression that show compression traces; and images with compression that show compression traces. Let C be the subset of images that have undergone compression, and let T_c be the subset of images that show compression traces. One can then use this convex combination $i' = \alpha i + (1-\alpha) i_c$ to obtain images in each of the 4 disjoint subsets that form I:

- $i, i' \in (\neg C \quad \& \quad \neg T_c)$: Images that have neither compression nor present com-

Figure 6.3: Illustration of morphisms over I for compression and the convex combination with respect to having compression and compression traces. The compression transformation (red) is non-invertible and can produce only images that have undergone compression, either with or without traces of compression (i_c and i'_c). The convex combination (purple), on the other hand, is able to produce all possible combinations of images, such that its co-domain is all I.



pression traces. When $\alpha = 1$ the convex transformation $i' = \alpha i + (1 - \alpha)i_c$ yields i. For values of $alpha$ sufficiently close to 1 the transformation will map $(\neg C \ \& \ \neg T_c) \rightarrow (\neg C \ \& \ \neg T_c)$;

- $i_c, i' \in (C \ \& \ T_c)$: Images that have compression and present compression traces. Conversely, as α approaches 0, i' becomes closer to i_c. If $i_c \in (C \ \& \ T_c)$, when $i' = i_c$ then $i' \in (C \ \& \ T_c)$, which means there is at least one instance where $(\neg C \ \& \ \neg T_c) \rightarrow (C \ \& \ T_c)$;

- $i \in (\neg C \ \& \ \neg T_c)$, $i_c \in (C \ \& \ T_c)$, $i' \in (\neg C \ \& \ T_c)$: Images that do not have compression and present compression traces. The true positive results from Section 3.4.2 prove that this set exists for certain $0 < \alpha < 1$, then $(\neg C \ \& \ \neg T_c) \rightarrow (\neg C \ \& \ T_c)$;

- $i'_c, i' \in (C \ \& \ \neg T_c)$: Images that have compression and do not present com-

pression traces. This is achieved for $\alpha = 0$ if instead of i_c the linear combination is done with another image, i'_c, which has compression but presents no traces of compression. We know experimentally that such images exists (when compression factor is sufficiently low, traces might not be detectable). Therefore, the mapping $(\neg C \ \& \ \neg T_c) \rightarrow (C \ \& \ \neg T_c)$ also exists.

If $C = T_c$ (*i.e.*, all compressed images exhibit compression traces), a property could be perfectly mapped and identified. However, we know from experimentation that $C \neq T_c$ since the sets $(C \ \& \ \neg T_c)$ and $(\neg C \ \& \ T_c)$ are not empty.

For a machine learning or statistical approach to perfectly map the ideal definition of a property P in a subset T_p (as in contain traces of p), it has to achieve 100% accuracy over all I. This is something very hard to achieve, and at the same time unfair to ask of any machine learning technique or mathematical model. If our goal is that the symmetric difference between the two sets is empty ($P \ominus T_p = \emptyset$), both P and T_p should be defined in ways that can minimize this difference. In other words, if the property of "undergone compression" can be redefined in terms of "traces", it should provide for a more meaningful and consistent modelling. The Judge and Oracle example from Chapter 5 illustrate this: the judge described a high-level property of an image being real, or reliable as evidence in court (P), to which the Oracle was unable to find a matching infallible model based on traces (T_p) such that their symmetric difference was empty (*i.e.*, there are no ambiguous or undecidable cases).

6.2 The Chain of Justification

In a previous discussion in Chapter 5, I argued that the main limitations for solving the Judge's dilemma are based on communication and access to knowledge. Now, I have shown how this can be understood as hard to define sets P and T_p, such that $P \ominus T_p = \emptyset$. To clarify why, let us focus on the property from Fig. 6.3. The importance of C, or determining if an image has undergone compression is to form an argument, or a chain of **justification**. Consider the following causal chain of justification of an analyst evaluating an image:

R_0: The analyst has determined that the image has been tampered based on the assessment that both P_1 (some meaningful property for the analyst), and C (the compression property) are true;

R_{-1}: It is believed for C to be true because T_c is true;

R_{-2}: It is believed for T_c to be true because in the representation space of its mathematical model it was closer to the trained samples for which C was true, rather than to the ones for which C was false.

We use the notation R for reasoning, starting at R_0 (reason 0) and decreasing because at each step we have to go deeper into the past, building on previous justification. At R_{-2} we can visualize a fork: on one side we have to provide justification for why this closeness in representation space bears truth, and on the other we have to justify why we believe in the C or $\neg C$ status of our samples. Let us call them left and right threads, such that on R^l we justify the mathematical model, and on R^r the choice of samples. One can clearly see that more complex clauses will spawn new forks, requiring justification from even further sources. The R^l can be followed in such a way that eventually a mathematical proof is reached. However, it is not enough to have a mathematically correct and justified model on biased data, so R^r must have a proper rooting.

This sort of "causal chain" analysis is a methodological tool to survey beyond a single clause for justification. Different responses to the Gettier problems were formulated based on chains of justification (MCDERMID, 2007). The "No False Lemma" formulation, for instance, states the original proposition only constitutes knowledge if all propositions in its causal chain can be guaranteed to be true (CRAIG, 1999). This is, of course, a very strict definition that would be impossible to apply to DIF in practice, since a lot of information comes from unreliable sources. The main reason we discuss chains of justification is because they are an intuitive response to tackle justification issues in DIF. They naturally construct a narrative that could be used as expert testimony, and help us understand how we ground statements in the field. We follow this line of reasoning to explore its limitations in the context of DIF, in the next two subsections.

6.2.1 Samples and Unaccounted Statistical Features

When an analyst uses a sample-based forensic technique or model, (s)he is trusting the sample choices. (S)he trusts that there are no unwanted unobserved phenomena in the real scenario that differentiates it significantly from the experimental model setup. More specifically, (s)he trusts there are no unobserved phenomena that could void her/his justification for believing in the result of the technique. Now, this does not mean (s)he

trusts the technique has 100% accuracy, only that their choice of samples is statistically representative of the whole of images. I claim that this cannot be guaranteed given current DIF methodological practices.

As discussed in Chapter 3, an image takes form in a complex process where its statistical properties are shaped in different ways. Acquisition traces are the prime example of this, with CFA patterns being a property of the image almost completely separate from its visual content. An image can be filtered in a simple way to remove traces of CFA without changing the actual content in any noticeable way. This, however, alters its properties from a statistical point of view. But what about images that have no CFA traces because their photographic process did not involve a color filter array? Are they more similar to a filtered CFA-captured image due to to the fact it has no traces of CFA , or to the unfiltered version because a filtering operation changes it too drastically? It depends on the choice of metric used. Visually, they could be exactly the same, but what matters is how close they are in the representation space of the chosen models.

The reason that machine learning models produce statements that are hard to scrutinize is because of their dimensional complexity. It is almost impossible to avoid unobserved phenomena since the images on a training dataset might share hidden statistical similarities that will influence learning. Using different test and training sets help to reduce this bias, but what about properties that are ubiquitous enough that affect multiple datasets? Images captured with a Bayer-like CFA are arguably the most common today, but this could change soon. Convolutional neural network models have been trained to develop new, better filter patterns (HENZ; GASTAL; OLIVEIRA, 2018) that can have unpredictable effect on current techniques. Another concern is that image apps and cloud storage services can have their own proprietary algorithms for filtering and compressing images in different ways. Once these operations affect a large enough amount of images it might be impossible to avoid the differences in their statistical properties, and how they affect analysis. From our survey of the literature (Section 3.2), most research does not address this issue, or proposes ways to re-generalize its results. This could become problematic as the samples used for research become obsolete, poorly representing the diversity of images in the wild.

A key observations is that digital images are not natural phenomena. Because of this, one must be very careful when scientifically studying digital images. The information contained in images may, frequently, be modelled after natural phenomena (*i.e.*, photography), but it this is not a necessary condition. An experiment on DIF is not performed

to determine truth about reality, or a subset of reality. DIF relies on experimentation to test its instruments, and give them credibility. When a researcher tests a technique against a forensics dataset, his results are only as expressive as the dataset represents the reality of images in general. If the JPEG standard changes, or a new form of compression becomes the norm, available techniques (which might have resulted from years of research on forensics) might not be applicable to the new images. Not only techniques that were explicitly based on JPEG compression would become obsolete, but ones that had their performance evaluated on JPEG images could be put in question.

6.2.2 Circular Justification, Precedence, and Bayes

It may seem that causal chains have a tree structure, as each justification step can branch into many, but loops can easily arise, transforming it into a graph. Clause R_{-1}, for instance, refers to a mapping between a high level property and traces, as discussed in Section 6.1.1. When considering this dualism in reference a loop arises in the form of a "chicken and egg" situation. What comes before, C, or T_c? The property C can be used and referenced in high-level because at one point T_c could be shown to be measured with some degree of certainty. However, as demonstrated, they are disjoint properties. In other words, if there existed no reliable technique to measure T_c, the analyst would not use C as justification.

In the clock example, the observer S looked at the traces of an object to determine it was a clock: round shaped, plastic frame, hanging on the wall, three handles and number markings from one to twelve. All these traces are evidence that it was a clock, and in fact one would have a hard time defending that such an object was not a clock. To justify the belief in that the time was 2:30, however, all these evidences were irrelevant. Being a clock is not a sufficient justification. To be able to tell the time correctly, a clock needs to be a working time-measuring device. Furthermore, if the clock is either delayed or advanced it cannot provide justified evidence for the time it is showing. From the point of view of S this is a dire dilemma, as one can only be sure that the clock is working and on time if (s)he knows precisely what time it is.

In our daily lives we avoid such dilemmas by developing intuition over accumulated experiences, so that we are not constantly questioning every event. If S frequently encounters stopped clocks and this causes her/him to be late often, this will reduce the trust in them, and in consequence, their use as justification. This is an interesting line of

reasoning, which has been explored by epistemologists and developed into a framework of Bayesian epistemology (BOVENS; HARTMANN, 2004). Through Bayesian statistics, it is possible to model reasoning processes for rational agents, where justification gives way to probabilistic masses. One does not need a justification for believing in p other than the belief that the probability of p is higher than $\neg p$.

Under a Bayesian Epistemologic framework one can envision a solution to the obsolescence of DIF research, as it provides mathematical tools for updating beliefs. On the other hand, as discussed in Section 5.6.1, one must be careful with the use of credences and abstracting people as purely rational agents (HORGAN, 2017). The more we base DIF on purely numerical models, even if they are consistent, the harder it is to provide applications that can be consumed by society. How should an analyst, for instance, explain to a jury the difference between a technique being 60% or 70% sure that an image contains traces of double compression? Without solid background knowledge, it is hard for someone to appropriately weight this information when making a decision.

6.3 The Rule Following Paradox

The problem that perhaps is the most influential in our discussion is the rule following paradox, firstly discussed by Wittgenstein (STERN, 2004), and then reformulated through Kripke's interpretation (FITCH, 2004). The paradox stems from the fact that no course of action can be determined by a rule, because any course of action can be made out to accord with the rule. In other words, it is impossible from an outside observer to determine if something is following an expected set of rules, or different rules that have so far been observed to produce the same results.

This paradox is deeply rooted in any situation involving language, behavior, and rules, and many problems presented in this Thesis can be thought of as instances of this paradox. In its classical form, it alludes to the impossibility of accessing someone else's (or something else's) internal states to guarantee they are using expected rules, or have the same meanings. But it is much more than that! Even if one would have access to such internal states, it is impossible to guarantee they will be understandable. Imagine being able to read the mind of a person who speaks Chinese while you do not speak it yourself. Chapter 5 revolved around conflicts of this sort, but we used a different line of arguing to arrive at them.

A fundamental question that has been raised several times in our discussion is the

difference between *real*, and *fake* images. One way to reformulate this question is to say that *fake images* are ones that have been formed by a set of "non-natural", or "unethical" rules. This is, in reality, the main paradigm DIF follows: identifying traces of unruly behavior in images. One tests for many different traces, asking questions about the image to determine if it follows expected rules. In our court example, this is precisely what the judge is aiming to find out. For him, the image could have been transformed in a variety of extravagant ways, as long as the contextual information he needs to use as evidence has not been affected. Imagine, for instance, that a picture of a suspect in a crowd was found in someone's Instagram, filtered to be more aesthetically pleasing. If key elements to the case such as clothing, height, etc. can be analyzed from it, the non-nativity of the image could be permitted. For a photography contest, however, altering the aesthetics of an image could be considered ill-will.

There is extensive literature on the issue, and many responses were proposed specially on the literature of Philosophy of Language. Attacks on this paradox generally aim to reformulate its premises, and show that in real life we unconsciously solve situations involving it in practical ways. This is also true for the DIF cases, where to some extent we have been able to solve conflicts, and develop frameworks to treat them. Kripke's so-called "skeptical" solution explains that we derive rules and meaning in contractual terms, in relation to other peers rather than absolutely (FITCH, 2004). People may have a slightly different meaning for what "cutlery" could entail, but the core concept of eating utensils is sufficiently diffused that speakers of English can agree. Even someone unfamiliar with forks and knifes can use it "meaningfully".

For DIF to effectively become a regulator of trust in communication, it must provide an epistemological structure to support such "meaning contracts", where people can agree on. This goes beyond defining a standard of terms and practices, it requires the participation of peers in developing practical uses. "Peers" in this sense does not refer only to analysts, researchers, and judges, but mostly users that will participate in communication with digital images. DIF theory and applications should be accessible and actively used as much as possible, and epistemological efforts are an important part of this process.

6.4 Summary

In this Chapter we presented challenges for pinpointing the nature of knowledge and justification in DIF. These challenges are not in any way fault of DIF itself, but stem

from deeper, fundamental questions in Philosophy. The dualism between C and T_c discussed in Section 6.1.1 is also reflection of a dualism in the field of epistemology itself (HÁJEK; LIN, 2017). Traditional and Bayesian epistemologies share different parts of a puzzle to understand knowledge and its role in decision making. Most technical research in DIF is naturally closer to the Bayesian approach, specially due to its instrumentation. While this modelling is ideal to capture the numerical complexity of digital images, it is limited in expressing subjective phenomena such as intent, accountability, and in capturing context. Since we advocate for DIF as a regulator in communication rather than a simple provider of technical solutions, these aspects cannot be overlooked. In *A Tale of Two Epistemologies* (HÁJEK; LIN, 2017) the author presents a parallel between both sides, highlighting promising recent work that is bridging this gap. It is evident that DIF could benefit from this developments to unify all its different theoretical backgrounds.

7 CONCLUSION

In this thesis, I discussed the current state of the field of Digital Image Forensics from different angles, many of which are not common in the DIF literature. The goal was to understand the challenges in the path of DIF to become a regulator for digital images as information currency. In fact, this discussion could be extended to any form of digital media used for communication, such as videos, or audio, which are essentially signals stored in digital format. In my representation of DIF as the encounter of *digital*, *image*, and *forensics* (Fig. 1.4) we use *image* in a much more general sense, as the pictorial element of human language. So far it is the least explored aspect of DIF in the literature, and probably the most complex, too.

In Chapter 2 we state that "Humans Are Easily Fooled", but it is not necessarily due to lack of cognitive capacity in any way. As our results showed, people not only mistake fake for real, but also real for fake. According to our anecdotal accounts (Section 2.2.4), subjects derived reasoning from contextual cues beyond the single image that was being evaluated, even if our experiment was designed to avoid this. Some justified their responses based on inferred relations between people depicted in the pictures, or recognizing traffic laws. Even if this line of reasoning lead them to a wrong answer, it is still an ingenuous strategy.

Our research comparing image composition techniques against forensics (Chapter 3) shows that it is much harder to fool a computational model than a human. At least in controlled scenarios, DIF's tools are powerful in scrutinizing the slightest inconsistencies to reveal editing. However, in our qualitative tests (Fig. 3.4) a lot of effort was put into understanding their outputs, and in *which ways* they are revealing editing. If those images were part of an actual analysis task with no ground truths for comparison, it would be almost impossible to identify those unexpected patterns and use them in reasoning.

The planning approach developed on Chapter 4 was envisioned to give technical support to users' ingenuous strategies. Using techniques and actions, one is able to suggest flexible plans for inspecting an image. The *Forensics Domain* (Fig. 4.3) acts as repository of knowledge that is maintained and updated in a collaborative effort by the community. Our difficulties in validating such a solution, both theoretically and practically evidenced the importance of that component beyond our idealized architecture. To be able to combine knowledge into a Forensics Domain one needs a meaningful, yet simple language to express it. What are the basic ontological elements of forensics, then?

Does the current DIF literature supports a universal language to describe its various concepts? Furthermore, who should be allowed editing power over the Forensics Domain? Our research showed that it is possible to treat DIF as an automated planning problem, and it has great potential, but its practical use depends on the answers to those questions.

In my vision, the current DIF trinity 1.4 greatly focus on the "digital", "digital images", and "digital forensics": the technical aspects. In Chapter 5 I present an argument for the inevitable limitation of such one-sided approach. My argument may have used a fantastical entity, the Oracle, for its exposition power, but one does not need to rely on it to prove this point. The fact that the general public is oblivious to the existence of DIF and to its methods is evident. Image composition, on the other hand, already established its presence in modern culture through many powerful, practical applications.

It is unreasonable, however, to criticize a technical field for its preference for the technical. More than that, it is unreasonable to ask that it provides solutions to problems that are far outside its reach, both in scope and capacity. In Chapter 6, I clarify what some of those problems are, for they are part of the central link that unites the DIF trinity. It is through understanding how each field approaches these problems that we can find common ground for **practical** solutions, even if theoretically they remain unsolvable behind paradoxes.

7.1 Future Work

This thesis has a clear message that multi-disciplinarity is key, and applications should be the goal. Things move very fast in today's world, and everything is connected in different, sometimes unexpected ways (DELEUZE; GUATTARI, 1987). Millions of digital images are exchanged on a daily basis, so any type of change in the status quo will have an immediate effect on society, economy, politics, etc. A single post from a teenager in a social network can make a company lose over a billion dollars in market value (BLOOMBERG, 2018). In such a scenario, it is fundamental that we develop a certain maturity in how we deal with digital media in society, and DIF is in a central position to herald this message.

There is great potential for research on DIF aligned with many fields, some of which are directly discussed in the text, such as Psychology, and Law. The field of Law is very fragmented, as political and regional elements shape its rules. Each country is in a different stage of assimilating and regulating digital images. However, this multiplicity

can be interesting for exploring different "ecosystems" of legislation and attitude towards digital images. As one of the main justifications for the existence of DIF, Law could help steering technical research for meaningful applications.

At least in the near future, many parallel approaches to deal with trust and sharing of digital images should emerge and exist concurrently. In practice, the need for authentication depends on the risk of the applications, as discussed in Section 5.1. Different commercial solutions might arise, compete, and evolve in this distributed "ecosystem". It is possible that blockchain technologies (SWAN, 2015) will provide means for organizing such decentralized trust networks. Companies could provide, for instance, a service for watermarking and authenticating images using a blockchain in a similar way HTTPS certificates work. Someone loading images from a news source on social media could guarantee it is signed from an authentication service, for instance.

Current technical research on DIF should increase its attention to reproducibility and accessibility. As discussed in Chapter 3, there is a great focus in developing novel forensics techniques, exploring different combinations of features and machine learning approaches. This is an important effort, but having an excessive amount of tools increases the complexity of the analysis task. Some authors have already started to focus on practical methodologies for Law applications (IULIANI, 2016; THEKKEKOODATHIL; VIJAYARAGAVAN, 2018), which is one of the most straight-forward ways of applying existing technology.

It is hard to predict the future of DIF or to endorse a particular line of research in favor of others. This thesis investigated many aspects of what DIF is and could be in relation to its subjects: analysts, researchers, social media users, and so on. At the moment, it stands in a position of high responsibility, but is far from achieving its full potential. We hope that our research and discussion stimulate new ideas and ways of thinking about the role of DIF beyond the **digital**, the **image**, and **forensics**.

REFERENCES

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. **Journal of Economic Perspectives**, v. 31, n. 2, p. 211–36, 2017.

ASGHAR, K.; HABIB, Z.; HUSSAIN, M. Copy-move and splicing image forgery detection and localization techniques: A review. **Australian Journal of Forensic Sciences**, 2016.

AVIDAN, S.; SHAMIR, A. Seam carving for content-aware image resizing. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 26, n. 3, jul. 2007. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/1276377.1276390>.

BELL, S.; BALA, K.; SNAVELY, N. Intrinsic images in the wild. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 33, n. 4, p. 159:1–159:12, jul. 2014. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/2601097.2601206>.

BERTALMIO, M. et al. Image inpainting. In: **Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques**. [S.l.: s.n.], 2000.

BHARATI, A. et al. Detecting facial retouching using supervised deep learning. **IEEE Transactions on Information Forensics and Security**, 2016.

BIANCHI, T.; PIVA, A. Image forgery localization via block-grained analysis of jpeg artifacts. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 7, n. 3, p. 1003–1017, 2012.

BIANCHI, T.; PIVA, A. Image forgery localization via block-grained analysis of jpeg artifacts. **Information Forensics and Security, IEEE Transactions on**, 2012.

BLOOMBERG. **In One Tweet, Kylie Jenner Wiped Out $1.3 Billion of Snap's Market Value**. 2018. Https://www.bloomberg.com/news/articles/2018-02-22/snap-royalty-kylie-jenner-erased-a-billion-dollars-in-one-tweet. Accessed: 2018-2-3.

BOVENS, L.; HARTMANN, S. **Bayesian Epistemology**. OUP Oxford, 2004. (Oxford scholarship online). ISBN 9780199269754. Available from Internet: <https://books.google.com.br/books?id=geg-rX_qICIC>.

BOYADZHIEV, I. et al. User-guided white balance for mixed lighting conditions. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 31, n. 6, p. 200:1–200:10, nov. 2012. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/2366145.2366219>.

CARROLL, R.; AGARWALA, A.; AGRAWALA, M. Image warps for artistic perspective manipulation. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 29, n. 4, p. 127:1–127:9, jul. 2010. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/1778765.1778864>.

CARROLL, R.; RAMAMOORTHI, R.; AGRAWALA, M. Illumination decomposition for material recoloring with consistent interreflections. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 30, n. 4, p. 43:1–43:10, jul. 2011. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/2010324.1964938>.

CARVALHO, T. et al. Illuminant-based transformed spaces for image forensics. **IEEE Transactions on Information Forensics and Security**, v. 11, n. 4, p. 720–733, 2016.

CARVALHO, T.; FARID, H.; KEE, E. Exposing photo manipulation from user-guided 3d lighting analysis. In: **IS&T/SPIE Electronic Imaging**. [S.l.: s.n.], 2015.

CARVALHO, T. et al. Exposing digital image forgeries by illumination color classification. **IEEE Transactions on Information Forensics and Security**, 2013.

CHAI, M. et al. Dynamic hair manipulation in images and videos. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 32, n. 4, p. 75:1–75:8, jul. 2013. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/2461912.2461990>.

CHEN, M. et al. Determining image origin and integrity using sensor noise. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 3, n. 1, p. 74–90, 2008.

CHEN, T. et al. 3-sweep: Extracting editable objects from a single photo. **ACM Transactions on Graphics (TOG)**, ACM, v. 32, n. 6, p. 195, 2013.

CHIERCHIA, G. et al. A bayesian-mrf approach for prnu-based image forgery detection. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 9, n. 4, p. 554–567, 2014.

CHO, H. et al. Bilateral texture filtering. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 4, p. 128, 2014.

CHU, X.; CHEN, Y.; LIU, K. J. R. Detectability of the order of operations: An information theoretic approach. **IEEE Transactions on Information Forensics and Security**, v. 11, n. 4, p. 823–836, 2016.

CHU, X.; CHEN, Y.; LIU, K. R. An information theoretic framework for order of operations forensics. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on**. [S.l.], 2016.

CHUANG, Y.-Y. et al. A bayesian approach to digital matting. In: **Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on**. [S.l.: s.n.], 2001.

CIESIELSKI, K. **Set Theory for the Working Mathematician**. [S.l.]: Cambridge University, 1997.

CONOTTER et al. Assessing the impact of image manipulation and image context on users' perceptions of deception. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Human Vision and Electronic Imaging XIX**. [S.l.], 2014. v. 9014, p. 90140Y.

CONOTTER, V.; BOATO, G.; FARID, H. Detecting photo manipulation on signs and billboards. In: IEEE. **Image Processing (ICIP), 2010 17th IEEE International Conference on**. [S.l.], 2010. p. 1741–1744.

COOPER, A. J. The electric network frequency (enf) as an aid to authenticating forensic digital audio recordings – an automated approach. In: **Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice**. [S.l.: s.n.], 2008.

COZZOLINO, D.; POGGI, G.; VERDOLIVA, L. Copy-move forgery detection based on patchmatch. In: **IEEE International Conference on Image Processing**. [S.l.: s.n.], 2014.

CRAIG, E. **Knowledge and the State of Nature**. Oxford University Press, 1999. Available from Internet: <https://doi.org/10.1093/0198238797.001.0001>.

DAISY, M.; TSCHUMPERLÉ, D.; LÉZORAY, O. A fast spatial patch blending algorithm for artefact reduction in pattern-based image inpainting. In: **SIGGRAPH Asia 2013 Technical Briefs**. [S.l.: s.n.], 2013.

DANG-NGUYEN, D.-T. **Discrimination of Computer Generated versus Natural Human Faces**. Thesis (PhD) — University of Trento, UNITN, 2014.

DARABI, S. et al. Image melding: Combining inconsistent images using patch-based synthesis. **ACM Trans. Graph.**, Citeseer, v. 31, n. 4, p. 82–1, 2012.

DELEUZE, G.; GUATTARI, F. **A Thousand Plateaus: Capitalism and Schizophrenia**. University of Minnesota Press, 1987. (Capitalism and schizophrenia). ISBN 9780816614028. Available from Internet: <https://books.google.com.br/books?id=C948Tsr72woC>.

DIAS, Z.; ROCHA, A.; GOLDENSTEIN, S. Image phylogeny by minimal spanning trees. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 7, n. 2, p. 774–788, 2012.

DING, M.; TONG, R.-F. Content-aware copying and pasting in images. **The Visual Computer**, Springer, v. 26, n. 6-8, p. 721–729, 2010.

DIRIK, A. E.; SENCAR, H. T.; MEMON, N. Analysis of seam-carving-based anonymization of images against prnu noise pattern-based source attribution. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 9, n. 12, p. 2277–2290, 2014.

ECHEVARRIA, J. I. et al. Computational simulation of alternative photographic processes. **Computer Graphics Forum (Proc. EGSR 2013)**, 2013.

ELGA, A. Self-locating belief and the sleeping beauty problem. **Analysis**, v. 60, n. 2, p. 143–147, 2000.

ENDO, Y. et al. Matting and compositing for fresnel reflection on wavy surfaces. **Computer Graphics Forum (Proc. of Eurographics Symposium on Rendering 2012)**, 2012.

FAN, S. et al. Human perception of visual realism for photo and computer-generated face images. **ACM Trans. Appl. Percept.**, ACM, New York, NY, USA, v. 11, n. 2, p. 7:1–7:21, jul. 2014. ISSN 1544-3558. Available from Internet: <http://doi.acm.org/10.1145/2620030>.

FAN, Z.; QUEIROZ, R. de. Identification of bitmap compression history: Jpeg detection and quantizer estimation. **Image Processing, IEEE Transactions on**, 2003.

FARBMAN, Z.; FATTAL, R.; LISCHINSKI, D. Convolution pyramids. **ACM Trans. Graph.**, v. 30, n. 6, p. 175–1, 2011.

FARBMAN, Z. et al. Coordinates for instant image cloning. **ACM Transactions on Graphics (TOG)**, ACM, v. 28, n. 3, p. 67, 2009.

FARID, H. Exposing digital forgeries from jpeg ghosts. **IEEE Trans. Information Forensics and Security**, v. 4, n. 1, p. 154–160, 2009.

FARID, H. Image forgery detection. **IEEE Signal processing magazine**, IEEE, v. 26, n. 2, p. 16–25, 2009.

FARID, H.; BRAVO, M. J. Image forensic analyses that elude the human visual system. v. 7541, p. 754106, 2010.

FARID, H.; BRAVO, M. J. Perceptual discrimination of computer generated and photographic faces. **Digital Investigation**, Elsevier, v. 8, n. 3-4, p. 226–235, 2012.

FATTAL, R. Dehazing using color-lines. In: . [S.l.]: ACM, 2014. v. 34, n. 1, p. 13.

FERRARA, P. et al. Image forgery localization via fine-grained analysis of cfa artifacts. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 7, n. 5, p. 1566–1577, 2012.

FERRARA, P. et al. Unsupervised fusion for forgery localization exploiting background information. In: **Multimedia Expo Workshops (ICMEW), 2015 IEEE International Conference on**. [S.l.: s.n.], 2015. p. 1–6.

FINLAYSON, G. D.; HORDLEY, S. D.; DREW, M. S. Removing shadows from images. In: SPRINGER. **European conference on computer vision**. [S.l.], 2002. p. 823–836.

FITCH, G. **Saul Kripke**. [S.l.]: MQUP, 2004. (Philosophy Now). ISBN 9780773581821.

FONTANI, M. et al. A framework for decision fusion in image forensics based on dempster-shafer theory of evidence. **IEEE Trans. Information Forensics and Security**, v. 8, n. 4, p. 593–607, 2013.

FORLAB: Multimedia Forensics Laboratory. Http://www.forlab.org/en/. Accessed: 2017-3-25.

FRAASSEN, B. V. **The Scientific Image**. Clarendon Press, 1980. (Clarendon Library of Logic and Philosophy). ISBN 9780198244271. Available from Internet: <https://books.google.com.br/books?id=VLz2F1zMr9QC>.

GASTAL, E. S.; OLIVEIRA, M. M. Shared sampling for real-time alpha matting. v. 29, n. 2, p. 575–584, 2010.

GASTAL, E. S.; OLIVEIRA, M. M. Adaptive manifolds for real-time high-dimensional filtering. **ACM Transactions on Graphics (TOG)**, ACM, v. 31, n. 4, p. 33, 2012.

GASTAL, E. S. L.; OLIVEIRA, M. M. High-order recursive filtering of non-uniformly sampled signals for image and video processing. **Computer Graphics Forum**, 2015.

GAUTHIER, I. et al. Expertise for cars and birds recruits brain areas involved in face recognition. **Nature neuroscience**, Nature Publishing Group, v. 3, n. 2, p. 191–197, 2000.

GETTIER, E. L. Is justified true belief knowledge? **Analysis**, v. 23, n. 6, p. 121–123, 1963.

GHADIYARAM, D.; BOVIK, A. C. Massive online crowdsourced study of subjective and objective picture quality. **IEEE Transactions on Image Processing**, v. 25, n. 1, p. 372–387, Jan 2016. ISSN 1057-7149.

GOLDFARB, A.; TUCKER, C. **Digital economics**. [S.l.]: National Bureau of Economic Research, 2017.

Gordon, D. et al. IQA: Visual Question Answering in Interactive Environments. **ArXiv e-prints**, dec. 2017.

GRYKA, M.; TERRY, M.; BROSTOW, G. J. Learning to remove soft shadows. **ACM Transactions on Graphics (TOG)**, ACM, v. 34, n. 5, p. 153, 2015.

GUO, R.; DAI, Q.; HOIEM, D. Single-image shadow detection and removal using paired regions. In: **Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2011. (CVPR '11). ISBN 978-1-4577-0394-2.

HACOHEN, Y. et al. Optimizing color consistency in photo collections. **ACM Transactions on Graphics (TOG)**, ACM, v. 32, n. 4, p. 38, 2013.

HÁJEK, A.; LIN, H. A tale of two epistemologies? **Res Philosophica**, v. 94, n. 2, p. 207–232, 2017.

HARTLEY, R. I.; ZISSERMAN, A. **Multiple View Geometry in Computer Vision**. Second. [S.l.: s.n.], 2004.

HASS, C. **JPEGsnoop**. 2017. Available from Internet: <http://www.impulseadventure.com/photo/jpeg-snoop.html>.

HELMERT, M. The fast downward planning system. **Journal of Artificial Intelligence Research**, 2006.

HENZ, B.; GASTAL, E. S. L.; OLIVEIRA, M. M. Deep joint design of color filter arrays and demosaicing. **Computer Graphics Forum**, 2018.

HOLMES, O.; BANKS, M. S.; FARID, H. Assessing and improving the identification of computer-generated portraits. **ACM Trans. Appl. Percept.**, ACM, New York, NY, USA, v. 13, n. 2, p. 7:1–7:12, feb. 2016. ISSN 1544-3558. Available from Internet: <http://doi.acm.org/10.1145/2871714>.

HORGAN, T. Troubles for bayesian formal epistemology. **Res Philosophica**, Philosophy Documentation Center, v. 94, n. 2, p. 233–255, 4 2017. ISSN 2168-9105.

HSU, E. et al. Light mixture estimation for spatially varying white balance. v. 27, n. 3, p. 70, 2008.

HUANG, H. et al. "mind the gap": tele-registration for structure-driven image completion. **ACM Trans. Graph.**, v. 32, n. 6, p. 174–1, 2013.

HUANG, H.; ZHANG, L.; ZHANG, H.-C. Repsnapping: Efficient image cutout for repeated scene elements. v. 30, n. 7, p. 2059–2066, 2011.

HUANG, J.-B. et al. Image completion using planar structure guidance. **ACM Transactions on graphics (TOG)**, ACM, v. 33, n. 4, p. 129, 2014.

HULLIN, M. et al. Physically-based real-time lens flare rendering. **ACM Transactions on Graphics (TOG)**, ACM, v. 30, n. 4, p. 108, 2011.

HáJEK, A. Interpretations of probability. In: **In The Stanford Encyclopedia of Philosophy (Zalta**. [S.l.: s.n.], 2003.

IEEE. **IEEE IFS-TC Image Forensics Challenge**. 2013. Http://ifc.recod.ic.unicamp.br/fc.website/index.py. Accessed: 2017-3-25.

IULIANI, M. **Image Forensics in the Wild**. Thesis (PhD) — Universita di Firenze, Universita di Perugia, INdAM consorziate nel CIAFM, 2016.

IULIANI, M.; FABBRI, G.; PIVA, A. Image splicing detection based on general perspective constraints. In: **Proceedings of the Information Forensics and Security (WIFS), 2015 IEEE International Workshop**. [S.l.: s.n.], 2015.

IULIANI, M. et al. Image counter-forensics based on feature injection. In: **Proc. SPIE 9028, Media Watermarking, Security, and Forensics 2014**. [S.l.: s.n.], 2014. v. 9028, p. 902810–902810–15.

JOSHI, N. et al. Personal photo enhancement using example images. **ACM Trans. Graph.**, v. 29, n. 2, p. 12–1, 2010.

KAPLAN, M. It's not what you know that counts. **Journal of Philosophy**, Journal of Philosophy Inc, v. 82, n. 7, p. 350–363, 1985.

KARSCH, K. et al. Rendering synthetic objects into legacy photographs. In: . [S.l.]: ACM, 2011. v. 30, n. 6, p. 157.

KARSCH, K. et al. Automatic scene inference for 3d object compositing. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 3, p. 32, 2014.

KAUFMANN, P. et al. Finite element image warping. **Comput. Graph. Forum**, 2013.

KE, Y. et al. Detection of seam carved image based on additional seam carving behavior. **International Journal of Signal Processing, Image Processing and Pattern Recognition**, 2016.

KEE, E.; FARID, H. Exposing digital forgeries from 3-d lighting environments. **IEEE International Workshop on Information Forensics and Security**, 2010.

KEE, E.; JOHNSON, M.; FARID, H. Digital image authentication from jpeg headers. **Information Forensics and Security, IEEE Transactions on**, 2011.

KEE, E.; O'BRIEN, J. F.; FARID, H. Exposing photo manipulation from shading and shadows. **ACM Trans. Graph.**, Citeseer, v. 33, n. 5, p. 165–1, 2014.

KHOLGADE, N. et al. 3d object manipulation in a single photograph using stock 3d models. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 4, p. 127, 2014.

KOPF, J. et al. Quality prediction for image completion. **ACM Transactions on Graphics (TOG)**, ACM, v. 31, n. 6, p. 131, 2012.

LAFFONT, P.-Y. et al. Coherent intrinsic images from photo collections. **ACM Transactions on Graphics**, v. 31, n. 6, 2012.

LAFFONT, P.-Y. et al. Transient attributes for high-level understanding and editing of outdoor scenes. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 4, p. 149, 2014.

LALONDE, J.-F.; EFROS, A. A. Using color compatibility for assessing image realism. In: IEEE. **Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on**. [S.l.], 2007. p. 1–8.

LEE, K.-T.; LUO, S.-J.; CHEN, B.-Y. Rephotography using image collections. **Computer Graphics Forum**, 2011.

LEE, S.; EISEMANN, E. Practical real-time lens-flare rendering. In: **Proceedings of the Eurographics Symposium on Rendering**. [S.l.: s.n.], 2013. (EGSR '13).

LI, B. et al. Revealing the trace of high-quality jpeg compression through quantization noise analysis. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 10, n. 3, p. 558–573, 2015.

LIAO, J. et al. Automating image morphing using structural similarity on a halfway domain. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 5, p. 168, 2014.

LIENG, H.; TOMPKIN, J.; KAUTZ, J. Interactive multi-perspective imagery from photos and videos. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2012. v. 31, n. 2pt1, p. 285–293.

LIU, L. et al. Blind image quality assessment by relative gradient statistics and adaboosting neural network. **Signal Processing: Image Communication**, v. 40, p. 1 – 15, 2016. ISSN 0923-5965. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0923596515001708>.

LIU, L.; JIN, Y.; WU, Q. Realtime Aesthetic Image Retargeting. In: **Computational Aesthetics in Graphics, Visualization, and Imaging**. [S.l.: s.n.], 2010. ISBN 978-3-905674-24-8.

LIU, Y. et al. AutoStyle: Automatic style transfer from image collections to users' images. **Computer Graphics Forum (Proc. Eurographics Symposium on Rendering)**, 2014.

LOPEZ-MORENO, J. et al. Compositing images through light source detection. **Computers & Graphics**, Elsevier, v. 34, n. 6, p. 698–707, 2010.

LUKAS, J.; FRIDRICH, J.; GOLJAN, M. Digital camera identification from sensor pattern noise. **Information Forensics and Security, IEEE Transactions on**, 2006.

LUO, W.; HUANG, J.; QIU, G. Jpeg error analysis and its applications to digital image forensics. **Information Forensics and Security, IEEE Transactions on**, 2010.

MAHMOUD, K. W.; AL-RUKAB, A. H. A. Moment based copy move forgery detection methods. **International Journal of Computer Science and Information Security (IJCSIS)**, 2016.

MATLAB. **version 7.13.0.564 (R2011b)**. Natick, Massachusetts: The MathWorks Inc., 2011.

MCDERMID, D. Beyond "justification": Dimensions of epistemic evaluation - by william p. alston. **Philosophical Books**, Blackwell Publishing Ltd, v. 48, n. 2, p. 175–177, 2007. ISSN 1468-0149. Available from Internet: <http://dx.doi.org/10.1111/j.1468-0149.2007.440_8.x>.

MCDERMOTT, D. et al. **PDDL - The Planning Domain Definition Language**. [S.l.], 1998.

MINUTES of the Thirty-First Meeting of the Board of Directors ofthe Optical Society of America, Incorporated. **J. Opt. Soc. Am.**, 1948.

MORTENSEN, E. N.; BARRETT, W. A. Intelligent scissors for image composition. In: **Proceedings of the 22Nd Annual Conference on Computer Graphics and Interactive Techniques**. [S.l.: s.n.], 1995. (SIGGRAPH '95). ISBN 0-89791-701-4.

NAU, D.; GHALLAB, M.; TRAVERSO, P. **Automated Planning: Theory & Practice**. [S.l.: s.n.], 2004. ISBN 1558608567.

NIGHTINGALE, S. J.; WADE, K. A.; WATSON, D. G. Can people identify original and manipulated photos of real-world scenes? **Cognitive Research: Principles and Implications**, v. 2, n. 1, p. 30, Jul 2017. ISSN 2365-7464. Available from Internet: <https://doi.org/10.1186/s41235-017-0067-2>.

OECD. **Skills Matter**. OECD Publishing, 2016. Available from Internet: </content/book/9789264258051-en>.

OSTROVSKY, Y.; CAVANAGH, P.; SINHA, P. Perceiving illumination inconsistencies in scenes. **Perception**, SAGE Publications Sage UK: London, England, v. 34, n. 11, p. 1301–1314, 2005.

PANOZZO, D.; WEBER, O.; SORKINE, O. Robust image retargeting via axis-aligned deformation. **Computer Graphics Forum**, v. 31, n. 2pt1, p. 229–236, 2012.

PARRY, Z. B. Digital manipulation and photographic evidence: defrauding the courts one thousand words at a time. **U. Ill. JL Tech. & Pol'y**, HeinOnline, p. 175, 2009.

PENG, B. et al. Optimized 3d lighting environment estimation for image forgery detection. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 12, n. 2, p. 479–494, 2017.

PENG, F.; LI, J.; LONG, M. Discriminating natural images and computer generated graphics based on compound fractal features. **Journal of Computational Information Systems**, v. 9, n. 13, p. 101–5108, 2013.

PENG, F.; ZHOU, D.-l. Discriminating natural images and computer generated graphics based on the impact of cfa interpolation on the correlation of prnu. **Digital Investigation**, Elsevier, v. 11, n. 2, p. 111–119, 2014.

PÉREZ, P.; GANGNET, M.; BLAKE, A. Poisson image editing. **ACM Transactions on graphics (TOG)**, ACM, v. 22, n. 3, p. 313–318, 2003.

PIVA, A. An overview on image forensics. **ISRN Signal Processing**, 2013.

RADEMACHER, P. et al. Measuring the perception of visual realism in images. In: **Proceedings of the 12th Eurographics Workshop on Rendering Techniques**. London, UK, UK: Springer-Verlag, 2001. p. 235–248. ISBN 3-211-83709-4. Available from Internet: <http://dl.acm.org/citation.cfm?id=647653.732279>.

RIESS, C.; ANGELOPOULOU, E. Scene illumination as an indicator of image manipulation. In: **Proceedings of the 12th international conference on Information hiding**. [S.l.: s.n.], 2010.

RITTERFELD, U.; CODY, M.; VORDERER, P. **Serious Games: Mechanisms and Effects**. Routledge, 2009. ISBN 9780415993692. Available from Internet: <http://books.google.com.br/books?id=GwPf7tbO5mgC>.

ROCHA, A. et al. Vision of the unseen: Current trends and challenges in digital image and video forensics. **ACM Comput. Surv.**, 2011.

SCHETINGER, V. et al. Image forgery detection confronts image composition. **Computers & Graphics**, v. 68, n. Supplement C, p. 152 – 163, 2017. ISSN 0097-8493. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0097849317301498>.

SCHETINGER, V. et al. Humans are easily fooled by digital images. **Computers & Graphics**, v. 68, n. Supplement C, p. 142 – 151, 2017. ISSN 0097-8493. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0097849317301450>.

SCHRAMM, W. **The process and effects of mass communication**. University of Illinois Press, 1954. Available from Internet: <https://books.google.com.br/books?id=yDAoAAAAMAAJ>.

SELVIN, S. On the monty hall problem (letter to the editor). **American Statistician**, v. 29, n. 3, p. 134, 1975.

SHANNON, C. E. A mathematical theory of communication. **SIGMOBILE Mob. Comput. Commun. Rev.**, ACM, New York, NY, USA, v. 5, n. 1, p. 3–55, jan. 2001. ISSN 1559-1662. Available from Internet: <http://doi.acm.org/10.1145/584091.584093>.

SHIH, Y. et al. Style transfer for headshot portraits. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 4, p. 148, 2014.

SIAPERA, E. **Understanding new media**. [S.l.]: Sage, 2017.

SINGH, G.; SINGH, K. Improved jpeg anti-forensics with better image visual quality and forensic undetectability. **Forensic Science International**, Elsevier BV, v. 277, p. 133–147, 2017.

SINHA, P. et al. Face recognition by humans: Nineteen results all computer vision researchers should know about. **Proceedings of the IEEE**, IEEE, v. 94, n. 11, p. 1948–1962, 2006.

SINHA, S. N. et al. Image-based rendering for scenes with reflections. **ACM Trans. Graph.**, Citeseer, v. 31, n. 4, p. 100–1, 2012.

STERN, D. G. **Wittgenstein's Philosophical Investigations: An Introduction**. [S.l.]: Cambridge University Press, 2004.

STOKES, M. et al. **A Standard Default Color Space for the Internet - sRGB**. 1996. Available from Internet: <http://www.w3.org/Graphics/Color/sRGB>.

SUNKAVALLI, K. et al. Multi-scale image harmonization. **ACM Transactions on Graphics (TOG)**, v. 29, n. 4, p. 125, 2010.

SUTTHIWAN, P. et al. Rake transform and edge statistics for image forgery detection. In: IEEE. **Multimedia and Expo (ICME), 2010 IEEE International Conference on**. [S.l.], 2010. p. 1463–1468.

SWAN, M. **Blockchain: Blueprint for a New Economy**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2015. ISBN 1491920491, 9781491920497.

TAO, M. W.; JOHNSON, M. K.; PARIS, S. Error-tolerant image compositing. In: **European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2010.

TAO, M. W.; MALIK, J.; RAMAMOORTHI, R. Sharpening out of focus images using high-frequency transfer. **Computer Graphics Forum (Eurographics 2013)**, 2013.

THEKKEKOODATHIL, V. T.; VIJAYARAGAVAN, P. K. A model of legal and procedural framework for cybercrime investigation in india using digital image forensics. In: DASH, S. S.; DAS, S.; PANIGRAHI, B. K. (Ed.). **International Conference on Intelligent Computing and Applications**. Singapore: Springer Singapore, 2018. p. 163–172.

TRIVEDI, K. **Probability and statistics with reliability, queuing, and computer science applications**. New York: Wiley, 2002. ISBN 0-471-33341-7.

VISHWANATH, D.; GIRSHICK, A. R.; BANKS, M. S. Why pictures look right when viewed from the wrong place. **Nature Neuroscience**, 2005.

WADE, K. A.; GREEN, S. L.; NASH, R. A. Can fabricated evidence induce false eyewitness testimony? **Applied Cognitive Psychology**, John Wiley & Sons, Ltd., v. 24, n. 7, p. 899–908, 2010. ISSN 1099-0720. Available from Internet: <http://dx.doi.org/10.1002/acp.1607>.

WANAT, R.; MANTIUK, R. K. Simulating and compensating changes in appearance between day and night vision. **ACM Transactions on Graphics (TOG)**, ACM, v. 33, n. 4, p. 147, 2014.

WATTANACHOTE, K. et al. Tamper detection of jpeg image due to seam modifications. **IEEE Transactions on Information Forensics and Security**, 2015.

WENG, Y. et al. Hair interpolation for portrait morphing. **Computer Grap**, 2013.

WU, H. et al. Resizing by symmetry-summarization. **ACM Transactions on Graphics (TOG)**, ACM, v. 29, n. 6, p. 159, 2010.

XUE, S. et al. Understanding and improving the realism of image composites. **ACM Transactions on Graphics (TOG)**, ACM, v. 31, n. 4, p. 84, 2012.

XUE, S. et al. Crowd sourcing memory colors for image enhancement. In: ACM. **ACM SIGGRAPH 2012 Talks**. [S.l.], 2012. p. 48.

YANG, J. et al. An effective method for detecting double jpeg compression with the same quantization matrix. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 9, n. 11, p. 1933–1942, 2014.

YIN, T. et al. Detecting seam carving based image resizing using local binary patterns. **Computers & Security**, Elsevier, v. 55, p. 130–141, 2015.

YÜCER, K. et al. Transfusive image manipulation. **ACM Transactions on Graphics (TOG)**, ACM, v. 31, n. 6, p. 176, 2012.

ZANDI, M.; MAHMOUDI-AZNAVEH, A.; TALEBPOUR, A. Iterative copy-move forgery detection based on a new interest point detector. **IEEE Transactions on Information Forensics and Security**, 2016.

ZHENG, J. et al. Exposing image forgery by detecting traces of feather operation. **Journal of Visual Languages & Computing**, 2015.

ZHENG, Y. et al. Interactive images: cuboid proxies for smart image manipulation. **ACM Trans. Graph.**, Citeseer, v. 31, n. 4, p. 99–1, 2012.

ZHOU, S. et al. Parametric reshaping of human bodies in images. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 29, n. 4, p. 126:1–126:10, jul. 2010. ISSN 0730-0301. Available from Internet: <http://doi.acm.org/10.1145/1778765.1778863>.