Forensic Science International 231 (2013) 178–189

Contents lists available at SciVerse ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Toward image phylogeny forests: Automatically recovering semantically similar image relationships

Zanoni Dias, Siome Goldenstein, Anderson Rocha *

Institute of Computing, University of Campinas, Campinas, SP 13083-852, Brazil

ARTICLE INFO

Article history:
Received 17 October 2012
Received in revised form 2 May 2013
Accepted 7 May 2013
Available online 4 June 2013

Keywords:
Image phylogeny
Phylogeny trees
Kinship analysis
Digital forensics

ABSTRACT

In the past few years, several near-duplicate detection methods appeared in the literature to identify the cohabiting versions of a given document online. Following this trend, there are some initial attempts to go beyond the detection task, and look into the structure of evolution within a set of related images overtime. In this paper, we aim at automatically identify the structure of relationships underlying the images, correctly reconstruct their past history and ancestry information, and group them in distinct trees of processing history. We introduce a new algorithm that automatically handles sets of images comprising different related images, and outputs the phylogeny trees (also known as a *forest*) associated with them. Image phylogeny algorithms have many applications such as finding the first image within a set posted online (useful for tracking copyright infringement perpetrators), hint at child pornography content creators, and narrowing down a list of suspects for online harassment using photographs.

1. Introduction

Most of us know the power of social media and its importance in making people more connected. Indeed, the new century's first decade has seen a vast rise of modern social media. What was formerly restricted to college campuses now easily achieve over millions of people with astonishing media uploading and sharing rates. For instance, Youtube claims that one hour of video is uploaded to their computers per second or more than 86 thousand hours of video per day (c.f., http://www.youtube.com/t/press_statistics). More importantly, Youtube has more than four billion video views per day which shows that content is not only being uploaded but also consumed.

Such shear amount of data has brought to us challenges never imagined before. For example, within such massive amount of data it is common that several documents are duplicates or near-duplicates of one another. While it is straightforward to find exact duplicates among available media, that does not hold true when media objects undergo small modifications. It is fairly common small changes to occur during the redistribution, usually without interfering on their semantic meaning. This is what we call *near-duplicate* media objects. These modifications can include, among others, A/D or D/A conversions, (de)-coding, transmission noise, and small editing/corrections such as brightness adjustments, or cropping.

The identification of near-duplicates has received particular attention during the past few years and [1–3] are just some examples in this area. However, a challenging task which has been vastly overlooked until recently, arises when we need to identify which document is the original within a set of digital related objects, and the structure of generation of each of them. In this case, we need to go beyond the identification of near-duplicate documents. We have to consider a population of multimedia objects as a whole to study the relationships among objects and their past history, as a result from the way they have been generated and manipulated overtime.

Although most of the changes related to near-duplicate multimedia objects are natural and not necessarily harmful, sometimes the distribution itself might cause copyright infringement or even represent a criminal action [4,5]. In some situations, the spreading pattern of an image or video can help companies to understand demographics and effectiveness of an ad campaign or a product. The identification of the original image posted online can help the analysis of a copyright infringement complaint. The original image is also the best candidate for a forensic authenticity analysis [6].

These scenarios motivated the dawning of a new research subfield called *Multimedia Phylogeny* [4], to investigate the history and evolutionary process of digital objects. In other words, we are looking for the structure of modifications of multimedia objects.

Solutions to problems in Multimedia Phylogeny have many applications:

* Corresponding author. Tel.: +55 19 3521 5854; fax: +55 19 3521 5838.
E-mail addresses: zanoni@ic.unicamp.br (Z. Dias), siome@ic.unicamp.br (S. Goldenstein), anderson@ic.unicamp.br, anderson.rocha@gmail.com (A. Rocha).

0379-0738/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.forsciint.2013.05.002

(a) security and law-enforcement (e.g., by narrowing down a list of suspects for online harassment);
(b) forensics (e.g., finding original documents within a set of related ones and allowing for more advanced document forensic analyses);
(c) copyright enforcement (e.g., traitor tracing without the requirement of active source control solutions such as watermarking or fingerprinting);
(d) news tracking services (e.g., document relationships can feed news tracking services with key elements for determining the opinion forming process across time and space [7,8]);
(e) content-based retrieval systems (e.g., showing similar photographs but of different photographers to a user without any metadata analysis).

Only recently there have been the first attempts to go beyond the near-duplicate identification problem to pinpoint the structure of relationships within a set of objects [9,4,10,8,7]. However, these early investigations are constrained to the case of image and video near-duplicates, which are related by a set of possible transformations – e.g., cropping, affine warping (considering as special cases resampling, rotation and translation), brightness/contrast adjustment and lossy compression. The main objective of such prior work was to identify the phylogeny tree associated with a set of near-duplicate images or videos.

In this paper, we go beyond prior work on Near Duplicate Images (NDI) and aim at finding the phylogeny trees within a set of Semantically Similar Images (SSI). Prior work in the literature have assumed the existence of relationships when analyzing a set of images. In this paper, without assuming the existence of relationships, we automatically find when images share a chain of processing history. For a better understanding, we formally define NDI and SSI documents in Section 2.

We expand upon state-of-the-art solutions [4,10,9] and present a new algorithm that automatically deals with sets of images from different sources, finding the different phylogeny trees.

We faced this problem for the first time while performing a real forensic analysis. On April 5th, 2009 [11], the Brazilian newspaper Folha de São Paulo, a major news player in Brazil, published an article about President Dilma Rousseff, back then the Brazilian Chief of Staff and a potential candidate for the 2010s presidential election (currently the president of that country). This article claimed that during her participation in the resistance to the Brazilian dictatorship, in the 60s, she engaged in violent or terrorist activities, such as armed robberies and kidnappings. To support this claim, the newspaper printed an alleged image of Secretary Rousseff's dossier from the internal files of the Repression Police (see Fig. 1), arguing it was obtained from the Public Archive of São Paulo, responsible for housing this collection of documents from that period of time.

Ms. Rousseff, who always declared herself as participant in a non-violent resistance movement, denied the allegations in the article and hired us to perform a forensic analysis of the image's authenticity. The newspaper never provided us the original printed image. Additionally, the image was virally widespread over the internet even before the newspaper chose to publish it – there were hundreds of copies in many different websites and blogs. Most of the copies were not exact but each one could have undergone additional image processing operations such as rescaling, cropping, and color adjustments.

That was the point where we identified that the literature lacked a robust approach for associating related images overtime, and the turning point for creating the multimedia phylogeny area. We needed a technique to answer questions

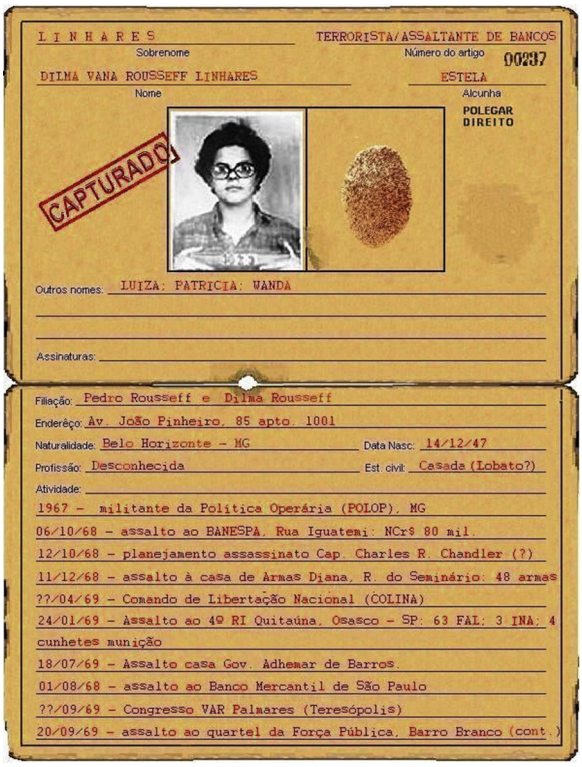

Fig. 1. The questioned object: the alleged image of Repression's Secret Police files on the Secretary of State Dilma Rousseff in 2009 as published by the Brazilian newspaper Folha de São Paulo.

such as: what image was the least modified one and, probably, the original released online (root of the tree)? What was the evolutionary associated tree? Could the tree's root (possibly the original image) be associated with auxiliary information of logs and website collected data to point out the perpetrator? In Section 6.4, we show the phylogeny tree associated with a few collected images related to this case and how we could put more effort analyzing the images on top of the tree instead of the leaves (which represent the least important modified versions for forensic purposes).

There are many forensic applications to image phylogeny solutions. Consider postings on the internet of private and/or abusive photographs of a regular person, such as in a bulling situation. Equally disturbing are online postings with fake and defamatory image content of celebrities or politicians (such as the case we discussed earlier). Similarly, our algorithm could help the fight against online child pornography (CP). Criminals use the internet to sell and disseminate CP images, and once these images go online, people usually redistribute them quickly using all sorts of hiding and changing methods.

Phylogeny algorithms can help us understand the evolutionary process among the set of replicated images. With the use of other metadata, and additional investigative work, it may be possible to track down the actual individuals who initially published the content online. Tools such as Microsoft's PhotoDNA [12,13] characterizes images using unique signatures looking for modified versions in an attempt to help law-enforcement to chase child pornographers. However, PhotoDNA normally looks only for exact or very similar copies, and it is not concerned about the evolutionary process among the images. Our solutions in image phylogeny look into the images' ancestry relationships, and is complementary to PhotoDNA.