LOYOLA UNIVERSITY CHICAGO 1870 AD MAJOREM DEI GLORIAM eCOMMONS

Loyola University Chicago
Loyola eCommons

Computer Science: Faculty Publications and Other Works

Faculty Publications and Other Works by Department

4-2-2022

Image Provenance Analysis

Daniel Moreira
Loyola University Chicago, dmoreira1@luc.edu

William Theisen
University of Notre Dame

Walter Scheirer
University of Notre Dame, walter.scheirer@nd.edu

Aparna Bharati
Lehigh University

Joel Brogan
Oak Ridge National Laboratory, broganjr@ornl.gov

See next page for additional authors

Follow this and additional works at: https://ecommons.luc.edu/cs_facpubs

 Part of the Computer Sciences Commons, and the Digital Communications and Networking Commons

Recommended Citation

Moreira, Daniel; Theisen, William; Scheirer, Walter; Bharati, Aparna; Brogan, Joel; and Rocha, Anderson. Image Provenance Analysis. Multimedia Forensics, , : 389-432, 2022. Retrieved from Loyola eCommons, Computer Science: Faculty Publications and Other Works, http://dx.doi.org/10.1007/978-981-16-7621-5_15

This Book Chapter is brought to you for free and open access by the Faculty Publications and Other Works by Department at Loyola eCommons. It has been accepted for inclusion in Computer Science: Faculty Publications and Other Works by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© The Author(s), 2022.

Authors

Daniel Moreira, William Theisen, Walter Scheirer, Aparna Bharati, Joel Brogan, and Anderson Rocha

This book chapter is available at Loyola eCommons: https://ecommons.luc.edu/cs_facpubs/336

Chapter 15
Image Provenance Analysis

Daniel Moreira, William Theisen, Walter Scheirer, Aparna Bharati, Joel Brogan, and Anderson Rocha

The literature of multimedia forensics is mainly dedicated to the analysis of single assets (such as sole image or video files), aiming at individually assessing their authenticity. Different from this, *image provenance analysis* is devoted to the joint examination of multiple assets, intending to ascertain their history of edits, by evaluating pairwise relationships. Each relationship, thus, expresses the probability of one asset giving rise to the other, through either global or local operations, such as data compression, resizing, color-space modifications, content blurring, and content splicing. The principled combination of these relationships unveils the provenance of the assets, also constituting an important forensic tool for authenticity verification. This chapter introduces the problem of provenance analysis, discussing its importance and delving into the state-of-the-art techniques to solve it.

15.1 The Problem

Consider a questioned media asset, namely a *query* (such as a digital image whose authenticity is suspect), and a large corpus of media assets (such as the Internet). Provenance analysis comprises the problem of (i) finding, within the available corpus,

D. Moreira · W. Theisen · W. Scheirer (✉)
University of Notre Dame, Notre Dame, IN, USA
e-mail: walter.scheirer@nd.edu

A. Bharati
Lehigh University, Bethlehem, PA, USA

J. Brogan
Oak Ridge National Laboratory, Oak Ridge, TN, USA

A. Rocha
University of Campinas, Campinas, Brazil

© The Author(s) 2022

H. T. Sencar et al. (eds.), *Multimedia Forensics*, Advances in Computer Vision and Pattern Recognition, https://doi.org/10.1007/978-981-16-7621-5_15

(a) (b) (c)

Fig. 15.1 Images that became viral on the web in the last decade, all with unknown sources. In **a**, the claimed world's first dab, supposedly captured during WWII. The dabbing soldier was highlighted, to make him more noticeable. In **b**, the claimed proof of an unlikely friendship between the Notorious B.I.G., on the left, and Kurt Cobain, on the right. In **c**, a photo of a supposed NATO meeting aimed at supporting a particular political narrative

(a) (b)

Fig. 15.2 A reverse image search of Fig. 15.1a leads to the retrieval of these two images, among others. Figure 15.1a is probably the result of cropping (**a**), which in turn is a color transformation of (**b**). The dabbing soldier was highlighted in **a**, to make him more noticeable. Image (b) is, thus, the source, comprising a behind-the-scenes picture from Dunkirk (2017)

the assets that directly and transitively share content with the query, as well as of (ii) establishing the derivation and content-donation processes that explain the existence of the query. Take, for example, the three queries depicted in Fig. 15.1, which became viral images in the last decade. Reasons for their virality range from the popularity of harmless pop-culture jokes and historical oddities (such as in the case of Fig. 15.1a, b) to interest in more critical political narratives and agendas (such as in Fig. 15.1c). Provenance analysis offers a principled and automated framework to debunk such media types by retrieving and associating other assets that help to elucidate their authenticity.

To get a glimpse of the expected outcome of performing provenance analysis, one could manually submit each one of the three queries depicted in Fig. 15.1 to a reverse image search engine, such as TinEye (2021), and try to select and associate the retrieved images to the queries by hand. This process is illustrated through Figs. 15.2, 15.3, and 15.4. Based on Fig. 15.2, for instance, one can figure out that the claimed world's first dab, supposedly captured during WWII, is a crop of Fig. 15.2a, which

(a) (b)

Fig. 15.3 A reverse image search of Fig. 15.1b leads to the retrieval of these two images, among others. Figure 15.1b is probably a composition of (**a**) and (**b**), where **a** donates the background, while **b** donates the Notorious B.I.G. on his car's seat. Cropping and color corrections are also performed to complete the forgery

(a) (b)

Fig. 15.4 A typical reverse image search of Fig. 15.1c leads to the retrieval of **a** but not **b**. Image **b** was obtained through one of the content retrieval techniques presented in Sect. 15.2 (context incorporation). Figure 15.1c is, thus, a composition of **a** and **b**, where **a** donates the background and **b** donates Vladimir Putin. Cropping and color corrections are also performed to complete the forgery

in turn is a color modified version of Fig. 15.2b, a well-known picture of the cast of the Hollywood movie Dunkirk (2017).

Figure 15.3, in turn, helps to reveal that Fig. 15.1b is actually a forgery (composition), where Fig. 15.3a probably serves as the background for the splicing of the Notorious B.I.G. on his car's seat, taken from Fig. 15.3a. In a similar fashion, Fig. 15.4 points out that Fig. 15.1c is also a composition, this time using Fig. 15.4a as background, and Fig. 15.4b as the donor of the portrayed individual. In this particular case, Fig. 15.4b is not easily found by a typical reverse image search. To do so, we had to perform *context incorporation*, a content retrieval strategy adapted to provenance analysis that is explained in Sect. 15.2.

In the era of misinformation and "fake news", there is a symptomatic crisis of trust in the media assets shared online. People are aware of editing software, with which even unskilled users can quickly fabricate and manipulate content. Although many of these manipulations have benign purposes (no, there is nothing wrong with the silly memes you share), some content is generated with malicious intent (general public deception and propaganda), and some modifications may undermine the ownership

of the media assets. Under this scenario, provenance analysis reveals itself as a convenient tool to expose the provenance of the assets, aiding in the verification of their authenticity, protecting their ownership, and restoring credibility.

In the face of the massive amount of data produced and shared online, though, there is no space for performing provenance analysis manually, such as formerly described. Besides the need for particular adaptations such as context incorporation (see Sect. 15.2), provenance analysis must also be performed at scale automatically and efficiently. By combining ideas from image processing, computer vision, graph theory, and multimedia forensics, provenance analysis constitutes an interesting interdisciplinary endeavor, into which we delve into in detail in the following sections.

15.1.1 The Provenance Framework

Provenance analysis can be executed at scale and fully automated by following a basic framework that involves two stages. Such a framework is depicted in Fig. 15.5. As one might observe, the first stage is always related to the activity of content retrieval, which incorporates a questioned media asset (*a.k.a.* the query) and a corpus of media assets to retrieve a selection of assets of interest that are related to the query.

Figure 15.6 depicts the expected outcome of content retrieval for Fig. 15.1b as the query. In the case of provenance analysis, the content retrieval activity must retrieve not only the objects that directly share content with the query but also transitively. Take, for example, images 1, 2, and 3 within Fig. 15.6, which all share some visual

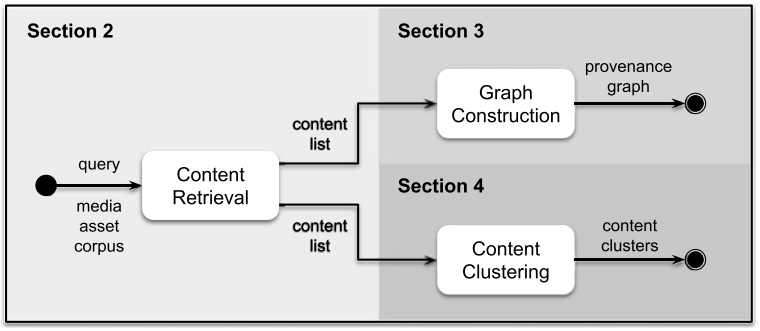


Fig. 15.5 The provenance framework. Provenance analysis is usually executed in two stages. Starting from a given query and a corpus of media assets, the first stage is always related to the content retrieval activity, which is herein explained in Sect. 15.2. The first stage's output is a list of assets of interest (content list), which is fed to the second stage. The second stage, in turn, may either comprise the activity of graph construction (discussed within Sect. 15.3) or the activity of content clustering (discussed within Sect. 15.4). While the former activity aims at organizing the retrieved assets in a provenance graph, the latter focuses on establishing meaningful asset clusters

Fig. 15.6 Content retrieval example for a given query. The desired output of content retrieval for provenance analysis is a list of media assets that share content with the query, either directly (such as objects 1, 2, and 3) or transitively (object 4, through object 1). Methods to perform content retrieval are discussed in Sect. 15.2

elements with the query. Image 4, however, has nothing in common with the query. But its retrieval is still desirable because it shares visual content with image 1 (the head of Tupac Shakur, on the right), hence it is related to the query transitively. Techniques to perform content retrieval for the provenance analysis of images are presented and discussed in Sect. 15.2.

Once a list of related media assets is available, the provenance framework's typical execution moves forward to the second stage, which has two alternate modes. The first, provenance graph construction, aims at computing the directed acyclic graph whose nodes individually represent the query and the related media assets and whose edges express the edit and content-donation story (e.g., cropping, blurring, splicing, etc.) between pairs of assets, linking seminal to derived elements. It, thus, embodies the provenance of the objects it contains. Figure 15.7 provides an example of provenance graph, constructed for the media assets depicted in Fig. 15.6. As shown, the query is probably a crop of image 1, which is a composition. It uses image 2 as background and splicing objects from images 3 and 4 to complete the forgery. Methods to construct provenance graphs from sets of images are presented in Sect. 15.3.

The provenance framework may be changed by replacing the second stage of graph construction with a content clustering approach. This setup is sound in the study of contemporary communication on the Internet, such as exchanging memes and the reproduction of viral movements. Dabbing (see Fig. 15.1a), for instance, is an example of this phenomenon. In these cases, the users' intent is not limited to retrieving near-duplicate variants or compositions that make use of a given query. They are also interested in the retrieval and grouping of semantically similar objects, which may greatly vary in appearance to elucidate the provenance of a trend. This situation is represented in Fig. 15.8. Since graph construction may not be the best approach to organize semantically similar media assets obtained during content retrieval, content clustering reveals itself as an interesting option, as we discuss in Sect. 15.4.

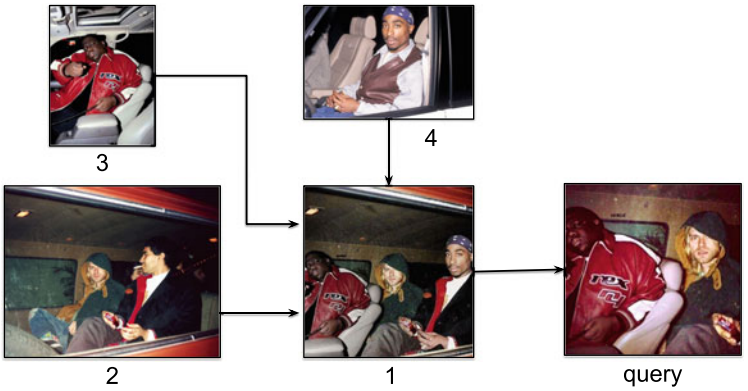


Fig. 15.7 Provenance graph example. This graph is constructed using the media assets retrieved in the example given through Fig. 15.6. In a nutshell, it expresses that the query is probably a crop of image 1, which in turn is a composition forged by combining the contents of images 2, 3, and 4. Methods to perform provenance graph construction are presented in Sect. 15.3

Fig. 15.8 Content retrieval of semantically similar assets. Some applications might aim at identifying particular behaviors or gestures that are depicted in a meme-style viral query. That might be the case for someone trying to understand the trend of dabbing, which is reproduced in all of the images above. In such a case, provenance content retrieval might still be helpful to fetch related objects. Graph construction, however, may be rendered useless due to the dominant presence of semantically similar objects rather than near-duplicates or compositions. In such a situation, we propose using content clustering as the second stage of the provenance framework, which we discuss in Sect. 15.4

15.1.2 Previous Work

Early Work: De Rosa et al. (2010) have mentioned the importance of considering groups of images instead of single images while performing media forensics. Starting with a set of images of interest, they have proposed to express pairwise image dependencies through the analysis of the mutual information between every pair of images. By combining these dependencies into a single correlation adjacency matrix for the entire image set, they have suggested the generation of a *dependency graph*,

whose edges should be individually evaluated as being in accordance with a particular set of image transformation assumptions. These assumptions should presume a known set of operations, such as rotation, scaling, and JPEG compression. Edges unfit to these assumptions should be removed.

Similarly, Kennedy and Chang (2008) have explored solutions to uncover the processes through which near-duplicate images have been copied or manipulated. By relying on the detection of a closed set of image manipulation operations (such as cropping, text overlaying, and color changing), they have proposed a system to construct *visual migration maps*: graph data structures devised to express the parent-child derivation operations between pairs of images, being equivalent to the dependency graphs proposed in De Rosa et al. (2010).

Image Phylogeny Trees: Rather than modeling an exhaustive set of possible operations between near-duplicate images, Dias et al. (2012) designed and adopted a robust image similarity function to compute a pairwise image similarity matrix **M**. For that, they have introduced an image similarity calculation protocol to generate **M**, which is widely used across the literature, including provenance analysis. This method is detailed in Sect. 15.3. To obtain a meaningful *image phylogeny tree* from **M**, to represent the evolution of the near-duplicate images of interest, they have introduced *oriented Kruskal*, a variation of Kruskal's algorithm that extracts an oriented optimum spanning tree from **M**. As expected, phylogeny trees are analogous to the aforementioned dependency graphs and visual migration maps.

In subsequent work, Dias et al. (2013) have reported a large set of experiments with their methodology in the face of a family of six possible image operations, namely scaling, warping, cropping, brightness and contrast changes, and lossy compression. Moreover, Dias et al. (2013) have also explored the replacement of oriented Kruskal with other phylogeny tree-building methods, such as *oriented Prim* and *Edmond's optimum branching*.

Melloni et al. (2014) have contributed to the topic of image phylogeny tree reconstruction by investigating ways to combine different image similarity metrics. Bestagini et al. (2016) have focused on the clues left by local image operations (such as object splicing, object removal, and logo insertion) to reconstruct the phylogeny trees of near-duplicates. More recently, Zhu and Shen (2019) have proposed heuristics to improve phylogeny trees by correcting local image inheritance relationship edges. Castelletto et al. (2020), in turn, have advanced the state of the art by training a denoising convolutional autoencoder that takes an image similarity adjacency matrix as input and returns an optimum spanning tree as the desired output.

Image Phylogeny Forests: All the techniques mentioned so far were conceived to handle near-duplicates. Aiming at providing a solution to deal with semantically similar images, Dias et al. (2013) have extended the oriented Kruskal method proposed in Dias et al. (2012) to what they named the *automatic oriented Kruskal*. This technique is an algorithm to compute a family of disjoint phylogeny trees (hence a phylogeny forest) from a given set of near-duplicate and semantically similar images, such that each disjoint tree describes the relationships of a particular group of near-duplicates.

In the same direction, Costa et al. (2014) have provided two extensions to the optimum branching algorithm proposed in Dias et al. (2013), namely *automatic optimum branching* and *extended automatic optimum branching*. Both solutions are based on the automatic calculation of branching cut-off execution points. Alternatively, Oikawa et al. (2015) have proposed the use of clustering techniques for finding the various disjoint phylogeny trees. Images coming from the same source (near-duplicates) should be placed in the same cluster, while semantically similar images should be placed in different clusters.

Milani et al. (2016), in turn, have suggested relying on the estimation of the geometric localization of captured viewpoints within the images as a manner to distinguish between near-duplicates (which should share viewpoints) from semantically similar objects (which should present different viewpoints). Lastly, Costa et al. (2017) have introduced solutions to improve the creation of the pairwise image similarity matrices, even in the presence of semantically similar images and regardless of the graph algorithm used to construct the phylogeny trees.

Multiple Parenting: Previously mentioned phylogeny work did not address the critical scenario of image compositions, in which objects from one image are spliced into another. Aiming at dealing with these cases, de Oliveira et al. (2015) have modeled every composition as the outcome of two parents (one *donor*, which provides the spliced object, and one *host*, which provides the composition background). Extended automatic optimum branching, proposed in Costa et al. (2014), should then be applied for the reconstruction of ideally three phylogeny trees: one for the near-duplicates of the donor, one for the near-duplicates of the host, and one for the near-duplicates of the composite.

Other Types of Media: Besides processing still images, some works in the literature have addressed the phylogeny reconstruction of assets belonging to other types of media, such as video (see Dias et al. 2011; Lameri et al. 2014; Costa et al. 2015, 2016; Milani et al. 2017) and even audio (see Verde et al. 2017). In particular, Oikawa et al. (2016) have investigated the role of similarity computation between digital objects (such as images, video, and audio) in multimedia phylogeny.

Provenance Analysis: The herein-mentioned literature of media phylogeny has made use of diverse metrics, individually focused on retrieving either the root, the leaves, or the ancestors of a node within a reconstructed phylogeny tree, evaluating the tree as a whole. Moreover, the datasets used in the experiments presented different types of limitations, such as either containing only images in JPEG format or lacking compositions with more than two sources (cf. object 1 depicted in Fig. 15.7).

Aware of these limitations and aiming to foster more research in the topic of multi-asset forensic analysis, the American Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST) have joined forces to introduce new terminology, metrics, and datasets (all herein presented, in the following sections), within the context of the Media Forensics (MediFor) project (see Turek 2021). Therefore, they coined the term *Provenance Analysis* to express a broader notion of phylogeny reconstruction, in the sense of

including not only the task of reconstructing the derivation stories of the assets of interest but also the fundamental step of retrieving these assets (as we discuss in Sect. 15.2). Furthermore, DARPA and NIST have suggested the use of directed acyclic graphs (*a.k.a. provenance graphs*), instead of groups of trees, to represent the derivation story of the assets better.

15.2 Content Retrieval

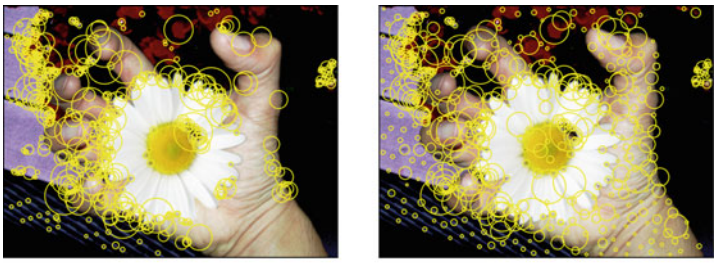
With the amount of content on the Internet being so vast, performing almost any type of computationally expensive analysis across the web's entirety or even smaller subsets for that matter is simply intractable. Therefore, when setting out to tackle the task of provenance analysis, a solution must start with an algorithm for retrieving a reasonably-sized subset of relevant data from a larger corpus of media. With this in mind, effective strategies for content retrieval become an integral module within the provenance analysis framework.

This section will focus specifically on *image* retrieval algorithms that provide results contingent on one or multiple images as queries into the system. This image retrieval is commonly known as reverse image search or more technically *Content-Based Image Retrieval* (CBIR). For provenance analysis, an appropriate CBIR algorithm should produce a corpus subset that contains a rich collection of images with relationships relevant to the provenance of the query image.

15.2.1 Approaches

Typical CBIR: Typical CBIR solutions employ multi-level representations of the processed images to reduce the semantic gap between the image pixel values (in the low level) and the system user's retrieval intent (in the highlevel, see Liu et al. 2007). Having provenance analysis in mind, the primary intent is to trace connections between images that mutually share visual content. For instance, when performing the query from Fig. 15.6, one would want a system to retrieve images that contain identical or near-identical structural elements (such as the corresponding images, respectively, depicting the Notorious B.I.G. and Kurt Cobain, both used to generate the composite query, but not other images of unrelated people sitting in cars). Considering that, we can describe an initial concrete example of a CBIR system that is not entirely suited to provenance analysis and further gradually provide methods to make it more suitable to the problem at hand.

A typical and moderately useful CBIR system for provenance analysis may rely on local features (*a.k.a.* keypoints) to obtain a low-level representation of the image content. Hence, it may consist of four steps, namely: (1) feature extraction, (2) feature compression, (3) feature retrieval, and (4) result ranking. Feature extraction comprises the task of computing thousands of n-dimensional representations for



(a) (b)

Fig. 15.9 SURF keypoints extracted from a given image. Each yellow circle represents a keypoint location, whose pixel values generate an n-dimensional feature vector. In **a**, a regular extraction of SURF keypoints, as defined by Bay et al. (2008). In **b**, a modified keypoint extraction dubbed distributed keypoints, proposed by Moreira et al. (2018), whose intention is also to describe homogeneous regions, such as the wrist skin and clapboard in the background. Images adapted from Daniel Moreira et al. (2018)

each image, ranging from classical handcrafted technologies, such as Scale Invariant Feature Transform (SIFT, see Lowe 2004) and Speeded-up Robust Features (SURF, see Bay et al. 2008), to neural network learned methods, such as Learned Invariant Feature Transform (LIFT, see Yi et al. 2016) and Deep Local Features (DELF, see Noh et al. 2017). Figure 15.9a depicts an example of SURF keypoints extracted from a target image.

Although the feature-based representation of images drastically reduces the amount of space needed to store their content, there is still a necessity for reducing their size, a task performed during the second CBIR step of feature compression. Take, for example, a CBIR system that extracts 1,000 64-dimensional floating-point SURF features from each image. Using 4 bytes for each dimension, each image occupies a total of $4 \times 64 \times 1000 = 256{,}000$ bytes (or 256 kB) of memory space. While that may seem relatively low from a single-image standpoint, consider an image database containing ten million images (which is far from being an unrealistic number, considering the scale of the Internet). This would mean the need for an image index on the order of 25 terabytes. Instead, we recommend utilizing Optimized Product Quantization (OPQ, see Ge et al. 2013) to reduce the size of each feature vector. In summary, OPQ learns grids of bins arranged along different axes within the feature vector space. These bin grids are rotated and scaled within the feature vector space to distribute optimally feature vectors extracted from an example training set. This provides a significant reduction in the number of bits required to describe a vector while keeping relatively high fidelity. For the sake of illustration, a simplified two-dimensional example of OPQ bins is provided in Fig. 15.10.

The third CBIR step (feature retrieval) aims at using the local features to index and compare, within the optimized n-dimensional space they constitute, and through Euclidean distance or similar method, pairs of image localities. Inverted File Indices (IVF, see Baeza-Yates and Ribeiro-Neto 1999) are the index structure commonly used in the feature retrieval step. Utilizing a feature vector binning scheme, such

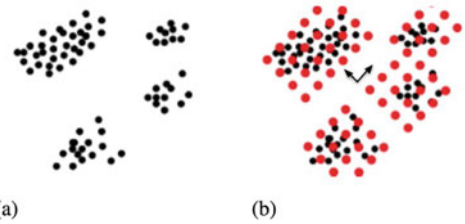


Fig. 15.10 A simplified 2D representation of OPQ. A set of points in the feature space **a** is re-described by rotating and translating grids of m bins to high-density areas (**b**), in red. The grid intervals and axis are learned separately for each dimension. Each dimension of a feature point can then be described using only m bits

as that of OPQ, an index is generated such that it stores a list of features contained within each bin. Each stored feature, in turn, points back to the image it was extracted from (hence the *inverted* terminology). The only calculation required is to determine the query feature's respective bin within the IVF to retrieve the nearest neighbors to a given query feature. Once this bin is known, the IVF can return the feature vectors' list in that bin and the nearest surrounding neighbor bins. Given that each feature vector points back to its source image, one can trace back the database images that are similar to the query. This simple yet powerful method provides easy scalability and distributability to index search. This is the main storage and retrieval structure used within powerful state-of-the-art search frameworks, such as the open-source Facebook Artificial Intelligence Similarity Search (FAISS) library introduced by Jeff Johnson et al. (2019).

Lastly, the result ranking step takes care of polling the feature-wise most similar database images to the query. The simplest metric with which one can rank the relatedness of the query image with the other images is feature voting (see Pinto et al. 2017). To perform it, one must iterate through each query feature and its retrieved nearest neighbor features. Then, by checking which database image each returned feature belongs to, one must accumulate a tally of how many features are matched to the query for each image. This final tally is then utilized as the votes for each database image, and these images are ranked (or ordered) accordingly. An example of this method can be seen in Fig. 15.11.

With these four steps implemented, one already has access to a rudimentary image search system capable of retrieving both near-duplicates and semantically similar images. While operational, this system is not particularly powerful when tasked with finding images with nuanced inter-image relationships relevant to provenance analysis. In particular, images that share only small amounts of content with the query image will not receive many votes in the matching process and may not be retrieved as a relevant result. Instead, this system can be utilized as a foundation for a retrieval system more fine-tuned to the task of image provenance analysis. In this regard, four methods to improve a typical CBIR solution are discussed in the following.

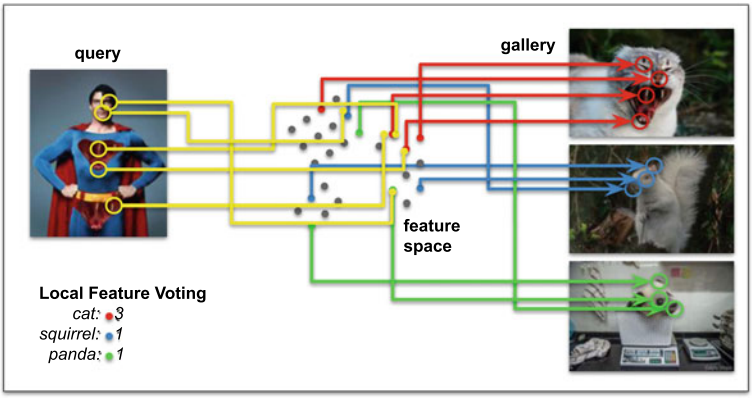


Fig. 15.11 A 2D simplified representation of IVF and feature voting. Local features are extracted from three images, representing a small gallery. The feature vector space is partitioned into a grid, and gallery features are indexed within this grid. Local features are then extracted from the query image of *Reddit Superman* (whose chest symbol and underwear are modified versions of the yawning cat's mouth depicted within the first item of the gallery). Each local query feature is fed to the index, and nearest neighbors are retrieved. Colored lines represent how each local feature matches the IVF features and subsequently map back to individual gallery features. We see how these feature matches can be used in a voting scheme to rank image similarities in the bottom left

Distributed Keypoints: To take the best advantage of these concepts, we must ensure that the extracted keypoints for the local features used within the indexing and retrieval pipeline do a good job describing the entire image. Most keypoint detectors return points that lie on corners, edges, or other high-entropy patches containing visual information. This, unfortunately, means that some areas within an image may be given too many vital points and may be over-described, while other areas may be entirely left out, receiving no keypoints at all. An area with no keypoints and, thus, no representation within the database image index has no chance of being correctly retrieved if a query with similar features comes along.

To mitigate over-description and under-description of image areas, Moreira et al. (2018) proposed an additional step in the keypoint detection process, which is the avoidance and removal of keypoints that present too much overlap with others. Such elements are then replaced with keypoints coming from weaker-entropy image regions, allowing for a more distributed content description. Figure 15.9b depicts an example of applying this strategy, in comparison with the regular SURF extraction approach.

Context Incorporation: One of the main reasons a typical CBIR solution performs poorly in an image provenance scenario is the nature behind many image manipulations within provenance cases. These manipulations often consist of composite images with small objects coming from donor images. An example of these types of relationships is shown in Fig. 15.12.

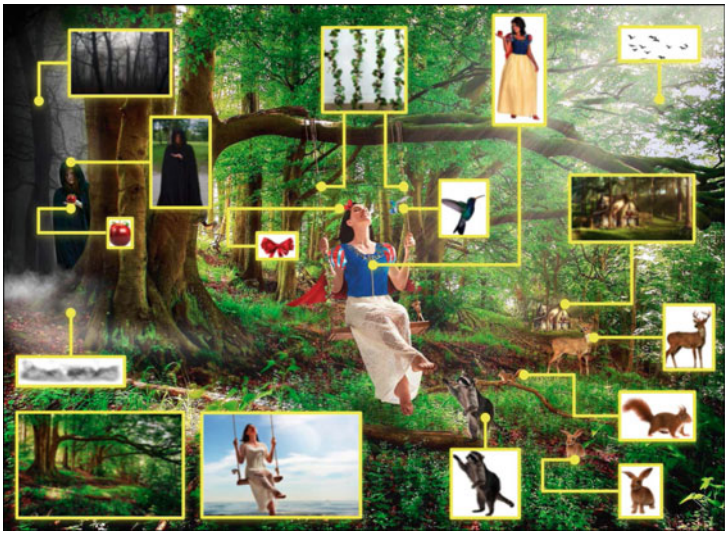


Fig. 15.12 An example of composite from the *r/photoshopbattles* subreddit. The relations of the composite with the images donating small objects, such as the hummingbird and the squirrel, challenge the retrieval capabilities of a typical CBIR solution. Context incorporation comes in handy in these situations. Image adapted from Joel Brogan et al. (2021)

These types of image relationships do not lend themselves to a naive feature voting strategy, as the size of the donated objects is often too small to garner enough local feature matches to impart a high vote score with the query image. We can augment the typical CBIR pipeline with an additional step, namely context incorporation, to solve this problem. Context incorporation takes into consideration the top N retrieved images for a given query, to accurately localize areas within the query and the retrieved images that differ from each other (most likely due to a composite or manipulation), to generate attention masks, and to re-extract features over only the distinct regions of the query, for a second retrieval execution. By using only these features, the idea is that the additional retrieval and voting steps will be driven towards finding the images that have donated small regions to the query due to the absence of distractors.

Figure 15.13 depicts an example where context incorporation is crucial to obtain the image that has donated Vladimir Putin (Fig. 15.13d) to a questioned query (Fig. 15.13a), in the first positions of the result rank. Different approaches for performing context incorporation, including attention mask generation, were proposed and benchmarked by Brogan et al. (2017), while an end-to-end CBIR pipeline employing such a strategy was discussed by Pinto et al. (2017).

Iterative Filtering: Another requirement of content retrieval within provenance analysis is the recovery of images directly related to the query and transitively related to it. That is the case of image 4 depicted within Fig. 15.6, which does not share content

Fig. 15.13 Context incorporation example. In **a**, the query image. In **b**, the most similar retrieved near-duplicate (top 1 image) through a typical CBIR solution. In **c**, the attention mask highlighting the different areas between **a** and **b**, after proper content registration. In **d**, the top 1 retrieved image after using **c** as a new query (a.k.a., context incorporation). The final result rank may be a combination of the two ranks after using **a** and **c** as a query, respectively. Image **d** is only properly retrieved, from the standpoint of provenance analysis, thanks to the execution of context incorporation

directly with the query, but is related to it through image 1. Indeed, any other near-duplicates of image 4 should ideally be retrieved by a flawless provenance-aware CBIR solution. Aiming at also retrieving transitively related content to the query, Moreira et al. (2018) introduced iterative filtering. After retrieving the first rank of images, the results are iteratively refined by suppressing near-duplicates of the query and promoting non-near-duplicates as new queries to the next retrieval iteration. This process is executed a number of times, leading to a set of image ranks for each iteration, which are then combined into a single one, at the end of the process.

Object-Level Retrieval: Despite the modifications above to improve typical CBIR solutions towards provenance analysis, local-feature-based retrieval systems do not inherently incorporate structural aspects into the description and matching process. For instance, any retrieved image with a high match vote score could still, in fact, be completely dissimilar to the query image. That happens because the matching process does not take into account the position of local features with respect to each other within an image. As a consequence, unwanted database images that contain features individually similar to the query's features, but in a different pattern, may still be ranked highly.

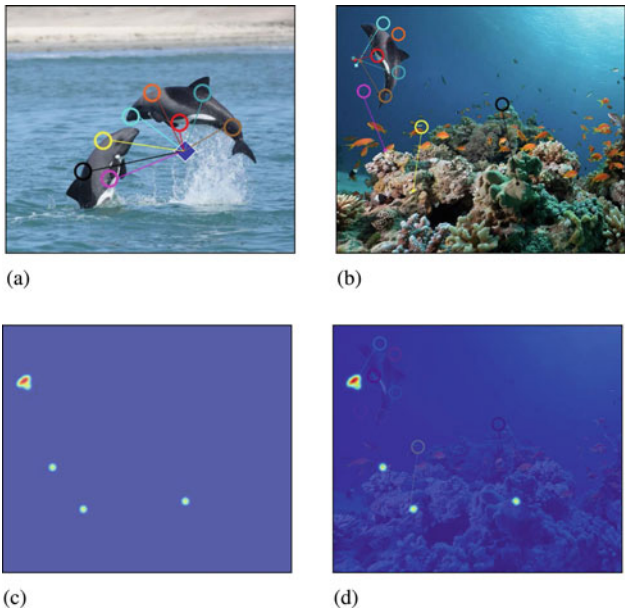


Fig. 15.14 An example of how OS2OS matching is accomplished. In **a**, a query image has the local feature keypoints mapped relative to a computed centroid location. In **b**, matching features from a database image are projected to an estimated centroid location using the vectors shown in **a**. In **c**, the subsequent vote accumulation matrix is created. In **d**, the accumulation matrix is overlayed on the database image. The red area containing many votes shows that the query and database images most likely share an object in that location

To avoid this behavior, Brogan et al. (2021) modified the retrieval pipeline to account for the structural layout of local features. To do so, they proposed a method to leverage the scale and orientation components calculated as part of the SURF keypoint extraction mechanism and to perform a transform estimation similar to the *generalized Hough voting* (see Ballard 1981), relative to a computed centroid. Because each feature's scale and orientation are known, for both the query's and database images' features, database features can be projected to the query keypoint layout space. Areas that contain similar structures accumulate votes in a given location on a Hough accumulator matrix. By clustering highly active accumulation areas, one is able to quickly determine local areas shared between images that have structural consistency with each other. A novel ranking algorithm then leverages these areas within the accumulator to subsequently score the relationship between images. This entire process is called "objects in scenes to objects in scenes" (OS2OS) matching. An example of how the voting accumulation works is depicted in Fig. 15.14.

15.2.2 Datasets and Evaluation

The literature for provenance analysis has reported results of content retrieval over two major datasets, namely the Nimble Challenge (NC17, 2017) and the Media Forensics Challenge (MFC18, 2018) datasets.

NC17 (2017): On the occasion of promoting the Nimble Challenge 2017, NIST released an image dataset containing a development partition (Dev1-Beta4) specifically curated to support both research tasks for provenance analysis. Namely, provenance content retrieval (named provenance filtering by the agency) and provenance graph construction. This partition contains 65 image queries and 11,040 images either related to the queries through provenance graphs, or completely unrelated material (named distractors). The provenance graphs were manually created by image edition experts and include operations such as splicing, removal, cropping, scaling, blurring, and color transformations. As a consequence, the partition offers content retrieval ground truth composed of 65 expected image ranks. Aiming to increase the challenge offered by this partition, Moreira et al. (2018) extended the set of distractors by adding one million unrelated images, which were randomly sampled from Eval-Ver1, another partition released by NIST as part of the 2017 challenge. We rely on this configuration to provide some results of content retrieval and explain how the different provenance CBIR add-ons explained in Sect. 15.2.1 contribute to solve the problem at hand.

MFC18 (2018): Similar to the 2017 challenge, NIST released another image dataset in 2018, with a partition (Eval-Ver1-Part1) also useful for provenance content retrieval. Used to officially evaluate the participants of the MediFor program (see Turek 2021), this set contains 3,300 query images and over one million images, including content related to the queries and distractors. Many of these queries are composites, with the expected content retrieval image ranks provided as ground truth. Moreover, this dataset also provides ground-truth annotations as to whether a related image contributes only a particular small object to the query (such as in the case of image 4 donating Tupac Shakur's head to image 1, within Fig. 15.7), instead of an entire large background. These cases are particularly helpful to assess the advantages of using the object-level retrieval approach presented in Sect. 15.2.1 in comparison to the other methods.

As suggested in the protocol introduced by NIST (2017), the metric used to evaluate the performance of a provenance content retrieval solution is the CBIR recall of the images belonging to the ground truth rank, at three specific cut-off points. Namely, (i) $R@50$ (i.e., the percentage of ground-truth expected images that are retrieved among the top 50 assets returned by the content retrieval solution), (ii) $R@100$ (i.e., the percentage of ground truth images retrieved among the top 100 assets returned by the solution), and (iii) $R@200$ (i.e., the percentage of ground truth images retrieved among the top 200 assets returned by the solution). Since the recall expresses the percentage of relevant images being effectively retrieved, the method delivering higher recall is considered better.

In the following section, results in terms of a recall are reported for the different solutions presented in Sect. 15.2.1, over the aforementioned datasets.

15.2.3 Results

Table 15.1 summarizes the results of provenance content retrieval reported by Moreira et al. (2018) over the NC17 dataset. It helps to put into perspective some of the techniques detailed in Sect. 15.2.1. As one might observe, by comparing rows 1 and 2 of Table 15.1, the simple modification of using more keypoints (from 2,000 to 5,000 features) to describe the images within the CBIR base module already provides a significant improvement on the system recall. Distributed Keypoints, in turn, improve the recall of larger ranks (for $R@100$ and $R@200$), while Iterative Filtering alone allows the system to reach an impressive recall of 90% of the expected images among the top 50 retrieved ones. At the time of their publication, Moreira et al. (2018) found that a combination of Distributed Keypoints and Iterative Filtering led to the best content retrieval solution, represented by the last row of Table 15.1.

More recently, Brogan et al. (2021) performed new experiments on the MFC18 dataset, this time aiming at evaluating the performance of their proposed OS2OS approach. Table 15.2 compares the results of the best solution previously identified by Moreira et al. (2018), in row 1, with the addition of OS2OS, in row 2, and

Table 15.1 Results of provenance content retrieval over the NC17 dataset. Reported here are the average recall values of 65 queries at the top 50 ($R@50$), top 100 ($R@100$), and top 200 ($R@200$) retrieved images. Provenance add-ons on top of the CBIR base module were presented in Sect. 15.2.1

CBIR Base	Provenance Add-ons	R@50	R@100	R@200	Source
2,000 SURF features, OPQ	Context Incorporation	71%	72%	74%	Daniel Moreira et al. (2018)
5,000 SURF features, OPQ	Context Incorporation	88%	88%	88%	Daniel Moreira et al. (2018)
5,000 SURF features, OPQ	Distributed Keypoints	88%	90%	90%	Daniel Moreira et al. (2018)
5,000 SURF features, OPQ	Iterative Filtering	90%	90%	92%	Daniel Moreira et al. (2018)
5,000 SURF features, OPQ	Distrib. Keypoints, Iterative Filtering	91%	91%	92%	Daniel Moreira et al. (2018)

Table 15.2 Results of provenance content retrieval over the MFC18 dataset. Reported here are the average recall values of 3,300 queries at the top 50 ($R@50$), top 100 ($R@100$), and top 200 ($R@200$) retrieved images. Provenance add-ons on top of the CBIR base module were presented in Sect. 15.2.1. OS2OS stands for "objects in scene to objects in scene", previously presented as object-level retrieval

CBIR Base	Provenance Add-ons	R@50	R@100	R@200	Source
5,000 SURF features, OPQ	Distrib. Keypoints, Iterative Filtering	77%	81%	82%	Joel Brogan et al. (2021)
5,000 SURF features, OPQ	Distrib. Keypoints, OS2OS	83%	83%	84%	Joel Brogan et al. (2021)
1,000 DELF features, OPQ	Iterative Filtering	87%	90%	91%	Joel Brogan et al. (2021)
1,000 DELF features, OPQ	OS2OS	91%	93%	95%	Joel Brogan et al. (2021)

replacement of the SURF features with DELF (see Noh et al. 2017) features, in rows 3 and 4, over the MFC18 dataset. OS2OS alone (compare rows 1 and 2) improves the system recall for all the three rank cut-off points. Also, the usage of only 1,000 DELF features (a data-driven image description approach that relies on learned attention models, in contrast to the 5,000 handcrafted SURF features) significantly improves the system recall (compare rows 1 and 3). In the end, the combination of a DELF-based CBIR system and OS2OS leads to the best content retrieval approach, whose recall values are shown in the last row of Table 15.2.

Lastly, as explained before, the MFC18 dataset offers a unique opportunity to understand how the content retrieval solutions work for the recovery of dataset images that eventually donated small objects to the queries at hand since it contains specific annotations in this regard. Table 15.3 summarizes the results obtained by Brogan et al. (2021) when evaluating this aspect. As expected, the usage of OS2OS greatly improves the recall of small-content donors, as it can be seen through rows 2 and 4 of Table 15.3. Overall, the best content retrieval approach (a combination of DELF features and OS2OS) is able to retrieve only 55% of the expected small-content donors among the top 200 retrieved assets (see the last row of Table 15.3). This result indicates that more work still needs to be done in this specific aspect: improving the provenance content retrieval of small-content donor images.

Table 15.3 Results of provenance content retrieval over the MFC18 dataset, with focus on the retrieval of images that donated small objects to the queries. Reported here are the average recall values of 3,300 queries at the top 50 ($R@50$), top 100 ($R@100$), and top 200 ($R@200$) retrieved images. Provenance add-ons on top of the CBIR base module were presented in Sect. 15.2.1. OS2OS stands for "objects in scene to objects in scene", previously presented as object-level retrieval

CBIR Base	Provenance Add-ons	R@50	R@100	R@200	Source
5,000 SURF features, OPQ	Distrib. Keypoints, Iterative Filtering	28%	34%	42%	Joel Brogan et al. (2021)
5,000 SURF features, OPQ	Distrib. Keypoints, OS2OS	45%	48%	52%	Joel Brogan et al. (2021)
1,000 DELF features, OPQ	Iterative Filtering	41%	45%	49%	Joel Brogan et al. (2021)
1,000 DELF features, OPQ	OS2OS	51%	54%	55%	Joel Brogan et al. (2021)

15.3 Graph Construction

A provenance graph depicts the story of edits and manipulations underwent by a media asset. This section focuses on the provenance graph of images, whose vertices individually represent the image variants and whose edges represent the direct pairwise image relationships. Depending on the transformations applied to one image to obtain another, the two connected images can share partial to full visual content. In the case of partial content sharing, the source images of the shared content are called the *donor images* (or simply donors), while the resultant manipulated image is called the *composite* image. In full-content sharing, we have near-duplicate variants when one image is created from another through a series of transformations such as cropping, blurring, and color changes. Once a set of related images is collected from the first stage of content retrieval (see Sect. 15.2), a fine-grained analysis of pairwise relationships is required to obtain the full provenance graph. This analysis involves two major steps, namely (1) image similarity computation and (2) graph building.

Similarity computation involves understanding the degree of similarity between two images. It is a fundamental task for any visual recognition problem. Image matching methods are at the core of vision-based applications, ranging from handcrafted approaches to modern deep-learning-based solutions. A matching method is a similarity (or dissimilarity) score that can be used for further decision-making and classification. For provenance analysis, computing pairwise image similarity helps distinguish between direct versus indirect relationships. A selection of a feasible set of pairwise relationships creates a provenance graph. To analyze the closest provenance match to an image in the provenance graph, pairwise matching is performed for all possible image pairs in the set of k retrieved images. The similarity scores are

then recorded in a matrix $\mathbf{M}$ of size $k \times k$ where each cell indexed $\mathbf{M}(i, j)$ represents the similarity between image I_i and image I_j.

Graph building, in turn, comprises the task of constructing the provenance graph after similarity computation. The matrix containing the similarity scores for all pairs of images involved in the provenance analysis for each case can be interpreted as an adjacency matrix. This implies that each similarity score in this matrix is the weight of an edge in a complete graph of k vertices. Extracting a provenance graph requires selecting a minimal set of edges that span the entire set of relevant images or vertices (this can be different from k). If the similarity measure used for the previous stage is symmetric, the final graph will be undirected, whereas an asymmetric measure of the similarity will lead to a directed provenance graph. The provenance cases considered in the literature, so far, are spanning trees. This implies that there are no cycles within graphs, and there is at most one path to get from one vertex to another.

15.3.1 Approaches

There are multiple aspects of a provenance graph. Vertices represent the different variants of an image or visual subject, pairwise relationships between images (i.e., undirected edges) represent atomic manipulations that led to the evolution of the manipulated image, and directions for these relationships provide more precise information about the change. Finally, the last details are the specific operations performed on one image to create the other. The fundamental task for an image-based provenance analysis is, thus, performing image comparison. This stage requires describing an image using a global or a set of local descriptors. Depending on the methods used for image description and matching, the similarity computation stage can create different types of adjacency weight matrices. The edge selection algorithm then depends on the nature of the computed image similarity. In the rest of this section, we present a series of six graph construction techniques that have been proposed in the literature and represent the current state of the art in image provenance analysis.

Undirected Graphs: A simple and yet-effective graph construction solution was proposed by Bharati et al. (2017). It takes the top k retrieved images for the given query and computes the similarity between the two elements of every image pair, including the query, through keypoint extraction, description, and matching. Keypoint extractors and descriptors, such as SIFT (see Lowe 2004) or SURF (see Bay et al. 2008), offer a manner to highlight the important regions within the images (such as corners and edges), and to describe their content in a way that is robust to several of the transformations manipulated images might have been through (such as scaling, rotating, and blurring.). The quantity of keypoint matches that are geometrically consistent with the others in the match set can act as an image similarity score for each image pair.

As depicted in Fig. 15.15, two images that share visual content will present more keypoint matches (see Fig. 15.15a) than the ones that have nothing in common (see

(a) (b)

Fig. 15.15 Examples of geometrically consistent keypoint matching. In **a**, the matching of keypoints over two images that share visual content. In **b**, the absence of matches between images that do not look alike. The number of matching keypoint pairs can be used to express the similarity between two images

Fig. 15.15b). Consequently, a symmetric pairwise image adjacency matrix can be built by simply using the number of keypoint matches between every image pair. Ultimately, a maximum spanning tree algorithm, such as Kruskal's (1956) or Prim's (1957), can be used to generate the final undirected image provenance graph.

Directed Graphs: The previously described method has the limitation of generating only symmetric adjacency matrices, therefore, not providing enough information to compute the direction of the provenance graphs' edges. As explained in Sect. 15.1, within the problem of provenance analysis, the direction of an edge within a provenance graph expresses the important information of which asset gives rise to the other.

Aiming to mitigate this limitation and inspired by the early work of Dias et al. (2012), Moreira et al. (2018) proposed an extension to the keypoint-based image similarity computation alternative. After finding the geometrically consistent keypoint matches for each pair of images (I_i, I_j), the obtained keypoints can be used for estimating the homography H_{ij} that guides the registration of image I_i onto image I_j, as well as the homography H_{ji} that analogously guides the registration of image I_j onto image I_i.

In the particular case of H_{ij}, after obtaining the transformation $T_j(I_i)$ of image I_i towards I_j, $T_j(I_i)$ and I_j are properly registered, with $T_j(I_i)$ presenting the same size of I_j and the matched keypoints relying on the same position. One can, thus, compute the bounding boxes that enclose all the matched keypoints within each image, obtaining two correspondent patches R_1, within $T_j(I_i)$, and R_2, within I_j. With the two aligned patches at hand, the distribution of the pixel values of R_1 can be matched to the distribution of R_2, before calculating the similarity (or dissimilarity) between them.

Considering that patches R_1 and R_2 have the same width W and height H after content registration, one possible method of patch dissimilarity computation is the pixel-wise mean squared error (MSE):

$$MSE(R_1, R_2) = \frac{\sum_w^W \sum_h^H (R_1(w,h) - R_2(w,h))^2}{H \times W}, \tag{15.1}$$

where $R_1(w, h) \in [0, 255]$ and $R_2(w, h) \in [0, 255]$ are the pixel values of R_1 and R_2 at position (w, h), respectively.

Alternatively to MSE, one can express the similarity between $R1$ and $R2$ as the mutual information (MI) between them. From the perspective of information theory, MI is the amount of information that one random variable contains about another. From the point of view of probability theory, it measures the statistical dependence of two random variables. In practical terms, assuming each random variable as, respectively, the aligned and color-corrected patches R_1 and R_2, the value of MI can be given by the entropy of discrete random variables:

$$MI(R_1, R_2) = \sum_{x \in R_1} \sum_{y \in R_2} p(x, y) \log \left(\frac{p(x, y)}{\sum_x p(x, y) \sum_y p(x, y)} \right), \tag{15.2}$$

where $x \in [0, 255]$ refers to the pixel values of R_1, and $y \in [0, 255]$ refers to the pixel values of R_2. The $p(x, y)$ value regards the joint probability distribution function of R_1 and R_2. As explained by Costa et al. (2017), it can be satisfactorily approximated by

$$p(x, y) = \frac{h(x, y)}{\sum_{x,y} h(x, y)}, \tag{15.3}$$

where $h(x, y)$ is the joint histogram that counts the number of occurrences for each possible value of the pair (x, y), evaluated on the corresponding pixels for both patches R_1 and R_2.

As a consequence of their respective natures, while MSE is inversely proportional to the two patches' similarity, MI is directly proportional. Aware of this, one can either use (i) the inverse of the MSE scores or (ii) the MI scores directly as the similarity elements s_{ij} within the pairwise image adjacency matrix $\mathbf{M}$, to represent the similarity between image I_j and the transformed version of image I_i towards I_j, namely $T_j(I_i)$.

The homography H_{ji} is calculated in an analogous way to H_{ij} with the difference that $T_i(I_j)$ is manipulated by transforming I_j towards I_i. Due to this, the size of the registered images, the format of the matched patches, and the matched color distributions will be different, leading to unique MSE (or MI) values for setting s_{ji}. Since $s_{ij} \neq s_{ji}$, the resulting similarity matrix $\mathbf{M}$ will be asymmetric. Figure 15.16 depicts this process.

Upon computing the full matrix, the assumption introduced by Dias et al. (2012) is that, in the case of $s_{ij} > s_{ji}$, it would be easier to transform image I_i towards image I_j, than the contrary (i.e., I_j towards I_i). Analogously, $s_{ij} < s_{ji}$ would mean the opposite. This information can, thus, be used for edge selection. The oriented Kruskal (2012) solution (with a preference for higher adjacency weights) would help construct the final provenance graph.

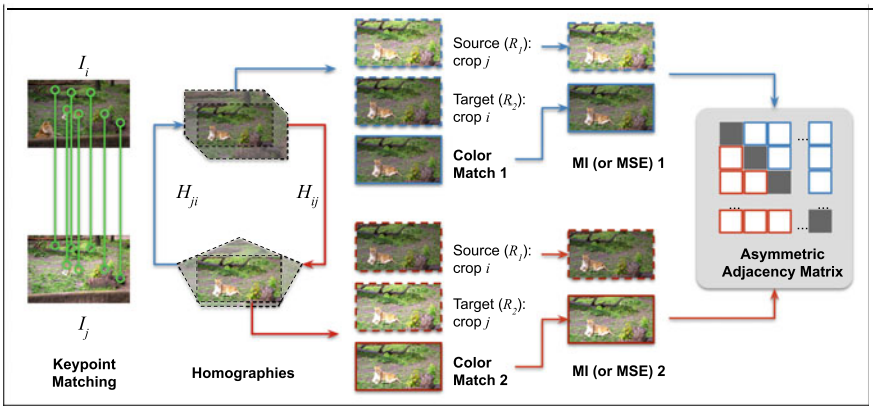


Fig. 15.16 Generation of an asymmetric pairwise image adjacency matrix. Based on the comparison of two distinct content transformations for each image pair (I_i, I_j), namely I_i towards I_j and vice-versa, this method allows the generation of an asymmetric pairwise image similarity matrix, which is useful for computing provenance graphs with directed edges

Clustered Graph Construction: As an alternative to the oriented Kruskal, Moreira et al. (2018) introduced a method of directed provenance graph building, which leverages both symmetric keypoint-based and asymmetric mutual-information-based image similarity matrices.

Inspired by Oikawa et al. (2015) and dubbed *clustered graph construction*, the idea behind such a solution is to group the available retrieved images in a way that only near-duplicates of a common image are added to the same cluster. Starting from the image query I_q as the initial expansion point, the remaining images are sorted according to the number of geometrically consistent matches shared with I_q, from the largest to the smallest. The solution then clusters probable near-duplicates around I_q, as long as they share *enough content*, which is decided based upon the number of keypoint matches (see Daniel Moreira et al. 2018). Once the query's cluster is finished (i.e., the remaining images do not share enough keypoint matches with the query), a new cluster is computed over the remaining unclustered images, taking another image of the query's cluster as the new expansion point. This process is repeated iteratively by trying different images as the expansion point until every image belongs to a near-duplicate cluster.

Once all images are clustered, it is time to establish the graph edges. Images belonging to the same cluster are sequentially connected into a single path without branches. This makes sense in scenarios containing sequential image edits where one near-duplicate is obtained on top of the other. As a consequence of the iterative execution and selection of different images as expansion points, the successful ones (i.e., the images that were helpful in the generation of new image clusters) fatally belong to more than one cluster, hence serving as graph bifurcation points. Orthogonal edges are established in such cases, allowing every near-duplicate image branch to be connected to the final provenance graph through an expansion point image as a

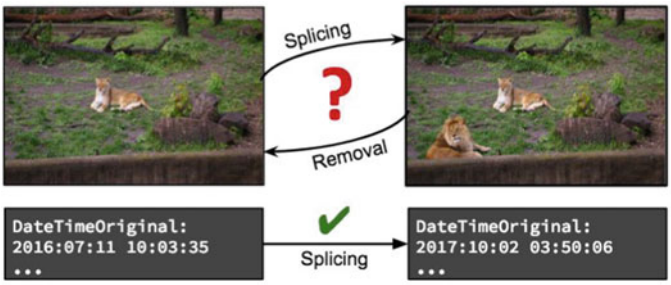


Fig. 15.17 Usage of metadata information to refine the direction of pairwise image provenance relationships. For the presented images, the executed manipulation operation could be either the splicing or the removal of the male lion. According to the image generation date metadata, the operation is revealed to be a splice since the image on the left is older. Adapted from Aparna Bharati et al. (2019)

joint. To determine the direction of every single edge, Moreira et al. (2018) suggested using the mutual information similarity asymmetry in the same way as depicted in Fig. 15.16.

Leveraging Metadata: Image comparison techniques may be limited depending upon the transformations involved in any given image's provenance analysis. In cases where the transformations are reversible or collapsible, the visual content analysis may not suffice for edge selection during graph building. Specifically, the homography estimation and color mapping steps involved in asymmetric matrix computation for edge direction inference could be noisy. To make this process more robust, it is pertinent to utilize other evidence sources to determine connections. As can be seen from the example in Fig. 15.17, it is difficult to point out the plausible direction of manipulation with visual correspondence, but auxiliary information related to the image, mostly accessible within the image files (*a.k.a.* image metadata), can increase confidence in predicting the directions.

Image metadata, when available, can provide additional evidence for directed edge inference. Bharati et al. (2019) identify highly relevant tags for the task. Specific tags that provide the time of image acquisition and editing, location, editing operation, etc. can be used for metadata analysis that corroborates visual evidence for provenance analysis. An asymmetric heuristic-based metadata comparison parallel to a symmetric visual comparison is proposed. The metadata comparison similarity scores are higher for image pairs (ordered) with consistency from more sets of metadata tags. The resulting visual adjacency matrix is used for edge selection, while the metadata-based comparison scores are used for edge direction inference. As explained in clustered graph construction, there are three parts of the graph building method, namely node cluster expansion, edge selection, and assigning directions to edges. The metadata information can supplement the last two depending on the specific stage at which it is incorporated. As metadata tags can be volatile in the world of intelligent forgeries, a conservative approach is to use them to improve the

confidence of the provenance graph obtained through visual analysis. The proposed design enables the usage of metadata when available and consistent.

Transformation-Aware Embeddings: While metadata analysis can improve the fidelity of edge directions in provenance graphs when available and not tampered with, local keypoint matching for visual correspondence faces challenges in image ordering for provenance analysis. Local matching is efficient and robust to finding shared regions between related images. This works well for connecting donors with composite images but can be insufficient in capturing subtle differences between near-duplicate images, which affect the ordering of long chains of operations. Establishing sequences of images that vary slightly based on the transformations requires differentiating between slightly modified versions of the same content.

Towards improving the reconstructions of globally-related image chains in provenance graphs, Bharati et al. (2021) proposed encoding awareness of the transformation sequence in the image comparison stage. Specifically, the devised method learns transformation-aware embeddings to better order related images in an edit sequence or provenance chain. The framework uses a patch-based siamese structure trained with an Edit Sequence Loss (ESL) using sets of four image patches. Each set is expressed as *quadruplets* or *edit sequences*, namely (i) the *anchor* patch, which represents the original content, (ii) the *positive* patch, a near-duplicate of the anchor after M image processing transformations, (iii) the *weak positive* patch, the positive patch after N transformations, and (iv) the *negative* patch, a patch that is unrelated to the others. The quadruplets of patches are obtained for training using a specific set of image transformations that are of interest to image forensics, particularly image phylogeny and provenance analysis, as suggested in Dias et al. (2012). For each anchor patch, random unit transformations are sequentially applied, one on top of the other's result, allowing to generate positive and weak positive patches from the anchor, after M and $M + N$ transformations, respectively. The framework aims at providing distance scores to pairs of patches, where the output score between the anchor and the positive patch is smaller than the one between the anchor and the weak positive, which, in turn, is smaller than the score between the anchor and the negative patch (as shown in Fig. 15.18).

Given a feature vector for an anchor image patch a, two transformed derivatives of the anchor patch p (positive) and p' (weak positive) where $p = T_M(a)$ and $p' = T_N(T_M(a))$, and an unrelated image patch from a different image n, ESL is a pairwise margin ranking loss computed as follows:

$$\begin{aligned} SL(a, p, p', n) = &\max(0, -y \times (d(a, p') - d(a, n)) + \mu_1) + \\ &\max(0, -y \times (d(p, p') - d(p, n)) + \mu_2) + \\ &\max(0, -y \times (d(a, p) - d(a, p')) + \mu_3) \end{aligned} \tag{15.4}$$

Here, y is the truth function which determines the rank order (see Rudin and Schapire 2009) and μ_1, μ_2, and μ_3 are margins corresponding to each pairwise distance term and are treated as hyperparameters. Both terms having the same sign

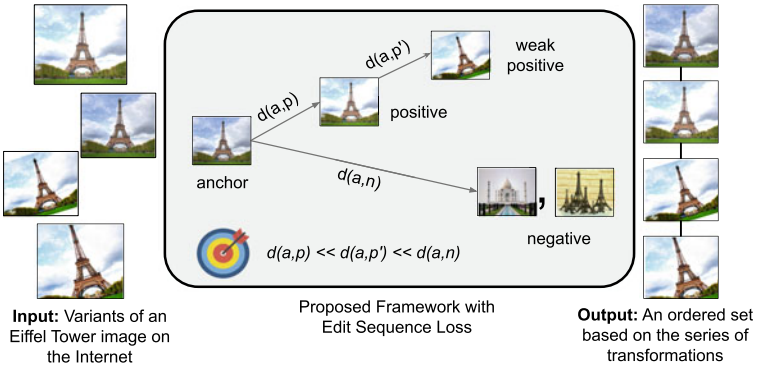


Fig. 15.18 Framework to learn transformation-aware embeddings in the context of ordering image sequences for provenance. Specifically, to satisfy an edit-based similarity precision constraint, *i.e.*, $d(a, p) < d(a, p') < d(a, n)$. Adapted from Aparna Bharati et al. (2021)

implies ordering is correct, and the loss is zero. A positive loss is accumulated when the ordering is wrong, and they are of opposite signs.

The above loss is optimized, and the model corresponding to the best measure for validation is used for feature extraction from patches of test images. Features learned with the proposed technique are used to provide pairwise image similarity scores. The value d_{ij} between images I_i and I_j is computed by matching the set of features (extracted from patches) from one image to the other using an iterative greedy brute-force matching strategy. At each iteration, the best match is selected as the pair of patches between image I_i and image I_j whose $l2$-distance is the smallest and whose patches did not participate in a match on previous iterations. This guarantees a deterministic behavior regardless of the order of the images, meaning that either comparing the patches of I_i against I_j or vice-versa will lead to the same consistent set of patch pairs. Once all patch pairs are selected, the average $l2$-distance is calculated and finally set as d_{ij}. The inverse of d_{ij} is then used to set both s_{ij} and s_{ji} within the pairwise image similarity matrix $\mathbf{M}$, which in this case is a symmetric one. Upon computing all values within $\mathbf{M}$ for all possible image pairs, a greedy algorithm (such as Kruskal's 1956) is employed to order these pairwise values and create an optimally connected undirected graph of images.

Leveraging Manipulation Detectors: A challenging aspect of image provenance analysis is establishing high-confidence direct relationships between images that share a small portion of content. Keypoint-based approaches may not suffice as there may not be enough keypoints in the shared regions, and global matching approaches may not appropriately capture the matching region's importance. To improve analysis of composite images where source images have only contributed a small region and determine the source image among a group of image variants, Zhang et al. (2020) proposed to combine a pairwise ancestor-offspring classifier with manipulation detection

approaches. They build the graph by combining edges based on both local feature matching and pixel similarity.

Their proposed algorithm attempts to balance global and local features and matching scores to boost performance. They start by using a weighted combination of the matched SIFT keypoints and the matched pixel values for image pairs that can be aligned, and null for the ones that cannot be aligned. A hierarchical clustering approach is used to group images coming from the same source together. For graph building within each determined cluster, the authors combine the likelihood of images being manipulated or extracted from a holistic image manipulation detector (see Zhang et al. 2020) and the pairwise ancestor score extracted by an L2-Net (see Tian et al. 2017). The image manipulation detector uses a patch-based convolutional neural network (CNN) to predict manipulations from a median-filtered residual image. For ambiguous cases where the integrity score may not be assigned accurately, a lightweight CNN-based ancestor-offspring network takes patch pairs as input and predicts one's scores to be derived from the other. The similarity scores used as edge weights are the average of the integrity and the ancestor scores from the two used networks. The image with the highest score among the smaller set of images is considered as the source. All incoming links to this vertex are removed to reduce confusion in directions. This one is then treated as the root of the arborescence built by applying Chu-Liu/Edmonds' algorithm (see Chu 1965; Edmonds 1967) on pairwise image similarities.

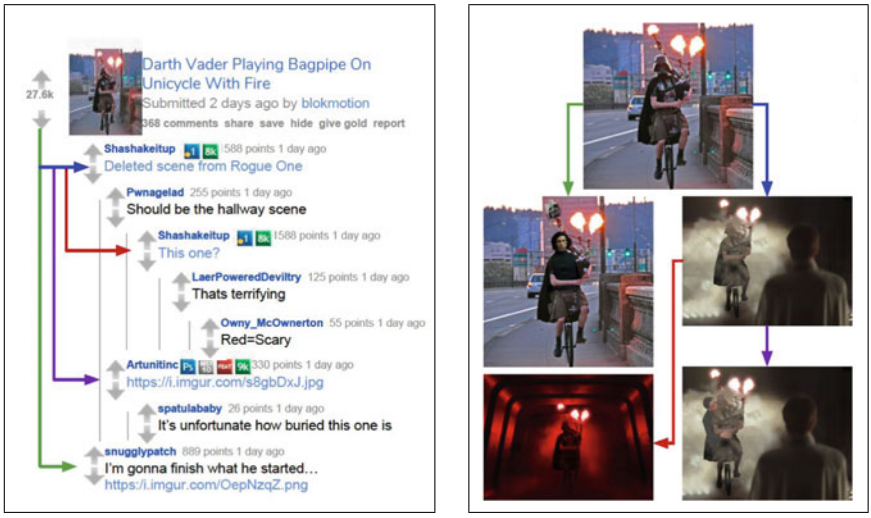
The different arborescences are connected by finding the best-matched image pair among the image clusters. If the matched keypoints are above a threshold, these images are connected, indicating a splicing or composition possibility. As reported in the following section, this method obtains state-of-the-art results on the NIST challenges (MFC18 2018 and MFC19 2019), and it significantly improves the computation of the edges of the provenance graphs over the Reddit Photoshop Battles dataset (see Brogan 2021).

15.3.2 Datasets and Evaluation

With respect to the step of provenance graph construction, four datasets stand out as publicly available and helpful benchmarks, namely NC17 (2017), MFC18 (2018) (both discussed in Sect. 15.2.2), MFC19 (2019), and the Reddit Photoshop Battles dataset (2021).

NC17 (2017): As mentioned in Sect. 15.2.2, this dataset contains an interesting development partition (Dev1-Beta4), which presents 65 image queries, each one belonging to a particular manually curated provenance graph. As expected, these provenance graphs are provided within the partition as ground truth. The number of images per provenance graph ranges from two to 81, with the average graph order being equal to 13.6 images.

MFC18 (2018): Besides providing images and ground truth for content retrieval (as explained in Sect. 15.2.2), the Eval-Ver1-Part1 partition of this dataset also pro-



(a) (b)

Fig. 15.19 A representation of how provenance graphs were obtained from the users' interactions within the Reddit r/photoshopbattles subreddit. In **a**, the parent-child posting structure of comments, which are used to generate the provenance graph depicted in **b**. Equivalent edges across the two images have the same color (either green, purple, blue, or red). Adapted from Daniel Moreira et al. (2018)

vides provenance graphs and 897 queries aiming at evaluating graph construction. For this dataset, the average graph order is 14.3 images, and the resolution of its images is larger, on an average, when compared to NC17. Moreover, its provenance cases encompass a larger set of applied image manipulations.

MFC19 (2019): A more recent edition of the NIST challenge released a larger set of provenance graphs. The Eval-Part1 partition within the MFC19 (2019) dataset has 1,027 image queries, and the average order of the provided ground truth provenance graphs is equal to 12.7 image vertices. In this group, the number of types of image manipulations used to generate the edges of the graphs was almost twice the number of MFC18.

Reddit Photoshop Battles (2021): Aiming at testing the image provenance analysis solutions over more realistic scenarios, Moreira et al. (2018) introduced the Reddit Photoshop Battles dataset. This dataset was collected from images posted to the Reddit community known as r/photoshopbattles (2012), where professional and amateur image manipulators share doctored images. Each "battle" starts with a teaser image posted by a user. Subsequent users post modifications of either the teaser or previously submitted manipulations in comments to the related posts. By using the underlying tree comment structure, Moreira et al. (2018) were able to infer and collect 184 provenance graphs, which together contain 10,421 original and composite images. Figure 15.19 illustrates this provenance graph inference process.

To evaluate the available graph construction solutions, two configurations are proposed by the NIST challenge (2017). In the first one, named the "oracle" scenario, there is a strong focus on the graph construction task. It assumes that a flawless content retrieval solution is available, thus, starting from the ground-truth content retrieval image ranks to build the provenance graphs, with neither missing images nor distractors. In the second one, named "end-to-end" scenario, content retrieval must be performed before graph construction, thus, delivering imperfect image ranks (with missing images or distractors) to the step of graph construction. We rely on both configurations and on the aforementioned datasets to report results of provenance graph construction, in the following section.

Metrics: As suggested by NIST (2017), given a provenance graph $G(V, E)$ generated by a solution whose performance we want to assess, we compute the $F1$-measure (i.e., the harmonic mean of precision and recall) of the (i) retrieved image vertices V and of the (ii) established edges E, when compared to the ground truth graph $G'(V', E')$, with its V' and E' homologous components. The first metric is named vertex overlap (VO) and the second one is named edge overlap (EO), respectively:

$$VO(G', G) = \quad 2 \times \frac{|V' \cap V|}{|V'| + |V|}, \tag{15.5}$$

$$EO(G', G) = \quad 2 \times \frac{|E' \cap E|}{|E'| + |E|}. \tag{15.6}$$

Moreover, we compute the vertex and edge overlap (VEO), which is the $F1$-measure of retrieving both vertices and edges simultaneously:

$$VEO(G', G) = \quad 2 \times \frac{|V' \cap V| + |E' \cap E|}{|V'| + |V| + |E'| + |E|}. \tag{15.7}$$

In a nutshell, these metrics aim at assessing the overlap between G and G'. The higher the values of VO, EO, and VEO, the better the performance of the solution. Finally, in the particular case of EO and VEO, when they are both assessed for an approach that does not generate directed graphs (such as Undirected Graphs and Transformation-Aware Embeddings, presented in Sect. 15.3.1), an edge within E is considered a hit (i.e., a correct edge) when there is a homologous edge within E' that connects equivalent image vertices, regardless of the edges' directions.

15.3.3 Results

Table 15.4 puts in perspective the different provenance graph construction approaches explained in Sect. 15.3.1, when executed over the NC17 dataset. The provided results

were all collected in oracle mode, hence the high values of VO (above 0.9), since there are neither distractors nor missing images in the rank lists used to build the provenance graphs. A comparison between rows 1 and 2 within this table shows the efficacy of leveraging image metadata as additional information to compute the edges of the provenance graphs. The values of EO (and VEO, consequently) have a significant increase (from 0.12 to 0.45, and from 0.55 to 0.70, respectively), when metadata is available. In addition, by comparing rows 3 and 4, one can observe the contribution of the data-driven Transformation-Aware Embeddings approach, in the scenario where only undirected graphs are being generated. In both cases, the generated edges have no direction by design, making their edge overlap conditions easier to be achieved (since the order of the vertices within the edges become irrelevant for the computation of EO and VEO, justifying their higher values when compared to rows 1 and 2). Nevertheless, contrary to the first two approaches, these solutions are not able to define which image gives rise to the other within the established provenance edges.

Table 15.5 compares the current state-of-the-art solution (Leveraging Manipulation Detectors by Xu Zhang et al. 2020) with the official NIST challenge participation results of the Purdue-Notre Dame team (2018), for both MFC18 and MFC19 datasets. In both cases, the reported results refer to the more realistic end-to-end scenario, where performers must execute content retrieval prior to building the provenance graphs. As a consequence, the image ranks fed to the graph construction step are noisy, since they contain both missing images and distractors. For all the reported cases, the image ranks had 50 images and presented an average $R@50$ of around 90% (i.e., nearly 10% of the needed images are missing). Moreover, nearly 35% of the images within the 50 available ones in a rank are distractors, on average. The

Table 15.4 Results of provenance graph construction over the NC17 dataset. Reported here are the average vertex overlap (VO), edge overlap (EO), and vertex and edge overlap (VEO) values of 65 queries. These experiments were executed in the "oracle" scenario, where the image ranks fed to the graph construction step are perfect (i.e., with neither distractors nor missing images)

Graph construction approach	VO	EO	VEO	Source
Clustered graph construction	0.93	0.12	0.55	Daniel Moreira et al. (2018)
Clust. Graph Const., Leveraging Metadata	0.93	0.45	0.70	Aparna Bharati et al. (2019)
Undirected Graphs	0.90	†0.65	†0.78	Aparna Bharati et al. (2021)
Transformation-Aware Embeddings	1.00	†0.68	†0.85	Aparna Bharati et al. (2021)

†Values collected over undirected edges

Table 15.5 Results of provenance graph construction over the MFC18 and MFC19 datasets. Reported here are the average vertex overlap (VO), edge overlap (EO), and vertex and edge overlap (VEO) values of 897 queries, in the case of MFC18, and of 1,027 queries, in the case of MFC19. These experiments were executed in the "end-to-end" scenario, thus, building graphs upon imperfect image ranks (i.e., with distractors or missing images)

Dataset	Graph Const. approach	VO	EO	VEO	Source
MFC18	Clustered Graph Construction	0.80	0.27	0.54	MFC19 (2019)
	Leveraging Manipulation Detectors	0.82	0.40	0.61	Xu Zhang et al. (2020)
MFC19	Clustered Graph Construction	0.70	0.30	0.52	MFC19 (2019)
	Leveraging Manipulation Detectors	0.85	0.42	0.65	Xu Zhang et al. (2020)

Table 15.6 Results of provenance graph construction over the Reddit Photoshop Battles dataset. Reported here are the average vertex overlap (VO), edge overlap (EO), and vertex and edge overlap (VEO) values of 184 queries. These experiments were executed in the "oracle" scenario, where the image ranks fed to the graph construction step are perfect. "N.R." stands for not-reported values

Graph construction approach	VO	EO	VEO	Source
Clustered Graph Construction	0.76	0.04	0.40	Daniel Moreira et al. (2018)
Clust. Graph Const., Leveraging Metadata	0.76	0.09	0.42	Aparna Bharati et al. (2019)
Leveraging Manipulation Detectors	N.R.	0.21	N.R.	Xu Zhang et al. (2020)

best solution (contained in rows 2 and 4 within Table 15.5) still delivers low values of EO when compared to VO, revealing an important limitation of the available approaches.

Table 15.6, in turn, reports results on the Reddit Photoshop Battles dataset. As one might observe, especially in terms of EO, this set is a more challenging one for the graph construction approaches, except for the state-of-the-art solution (Leveraging Manipulation Detectors by Xu Zhang et al. (2020). While methods that have worked fairly well on the NC17, MFC18, and MFC19 datasets drastically fail in the case of the Reddit dataset (see EO values below 0.10 in the case of rows 1 and 2), the state-of-the-art approach (in the last row of the table) more than doubles the results

of EO. Again, even with this improvement, increasing the values of EO within graph construction solutions is still an open problem that deserves attention from researchers.

15.4 Content Clustering

In the study of human communication on the Internet and the understanding of the provenance of trending assets, such as memes and other forms of viral content, the users' intent is mainly focused on the retrieval, selection, and organization of semantically similar objects, rather than the gathering of near-duplicate variants or compositions that are related to a query. Under this scenario, although the step of content retrieval may be useful, the step of graph construction loses its purpose, since the available content is preferably related through semantics (e.g., different people on diverse scenes doing the same action, such as the "dabbing" trend depicted on Fig. 15.8), greatly varying in appearance, making the techniques presented on Sect. 15.3 less suitable.

Humans can only process so much data at once—if too much is present, they begin to be overwhelmed. In order to help facilitate the human processing and perception of the retrieved content, Theisen et al. (2020) proposed a new stage to the provenance pipeline focused on image clustering. Clustering the images based on shared elements helps triage the massive amounts of data that the pipeline has grown to accommodate. Nobody can reasonably review several million images to find emerging trends and similarities, especially without ordering in the image collection. However, when these images are grouped based on shared elements using the provenance pipeline, the number of items a reviewer would have to look at can decrease by several magnitude orders.

15.4.1 Approach

Object-Level Content Indexing: From the content retrieval techniques discussed in Sect. 15.2, Theisen et al. (2020) recommended the use of OS2OS matching (see Brogan et al. 2021) to index the millions of images eventually available, due to two major reasons. Firstly, to obtain a fast and scalable content retrieval engine. Secondly, to benefit from the OS2OS capability of comparing images through either large and global content matching or through many small object-wise local matches.

Affinity Matrix Creation: In the task of analyzing a large corpus of assets shared on the web to understand the provenance of a trend, the definition of a query (i.e., a questioned asset) is not always as straightforward as it is in the image provenance analysis case. For example, the memes shared during the 2019 Indonesian elections and discussed in William Theisen et al. (2020). In such cases, a natural question to ask

would be which memes in the dataset one should use as the queries, for performing the first step of content retrieval.

Inspired by a simplification of iterative filtering (see Sect. 15.2.1), Theisen et al. (2020) identified the cheapest option as being randomly sampling images from the dataset and iteratively using them as queries, for executing content retrieval until all the dataset images (or a sufficient number of them) are "touched" (i.e., they are retrieved by the content retrieval engine). There are several advantages to this, other than being easy to implement. Randomly sampling means that end-users would need to have no prior knowledge of the dataset and potential trends they are looking for. The cluster created at the end of the process would show "emergent" trends, which could even surprise the reviewer.

On the other hand, from a more forensics-driven perspective, it is straightforward to imagine a system in which "informed queries" are used. If the users already suspect that several specific trends may exist in the data, cropped images of the objects pertaining to a trend may be used as a query, thus, prioritizing the content that the reviewers are looking for. This might be a demanding process because the user must already have some sort of idea of the landscape of the data and must produce the query images themselves.

Following the suit in the solution proposed in William Theisen et al. (2020), prior to performing clustering, a particular type of image pairwise adjacency matrix (dubbed *affinity matrix*) must first be generated. By leveraging the provenance pipeline's pre-existing steps, this matrix can be constructed based on the retrieval of the many selected queries. To prepare for the clustering step, a number of queries need to be run through the content retrieval system, the number of queries, and the recall of them depending on what type of graph the user wants to model for the analyzed dataset. Using the retrieval step's output, a number of query-wise "hub" nodes are naturally generated, each of them having many connections. Consider, for example, that the user has decided to have a recall of the top 100 best matches for any selected query. This means that for every query submitted, there are 100 connections to other assets. These connections link the query image to each of the 100 matches, thus, imposing a "hubness" onto the query image. For the sake of illustration, Fig. 15.20 shows an example of this process, for the case of memes shared during the 2019 Indonesian elections.

By varying the recall and number of queries, the affinity matrix can be generated, and an order can be imposed. Lowering the recall will result in a decrease in hub-like nodes in the structure but will require more queries to be run to connect all the available images. The converse is also true. Once enough queries have been run, the final step of clustering can be executed.

Spectral Clustering: After the affinity matrix has been generated, Theisen et al. (2020) suggested the use of multiclass spectral clustering (see Stella and Shi 2003) to organize the available images. In the case of memes, this step has the effect of assigning each image to a hypothesized *meme genre*, similar to the definition presented by Shifman (2014). The most important and perhaps trickiest part is deciding on a number of clusters to create. While more clusters may allow for a more targeted

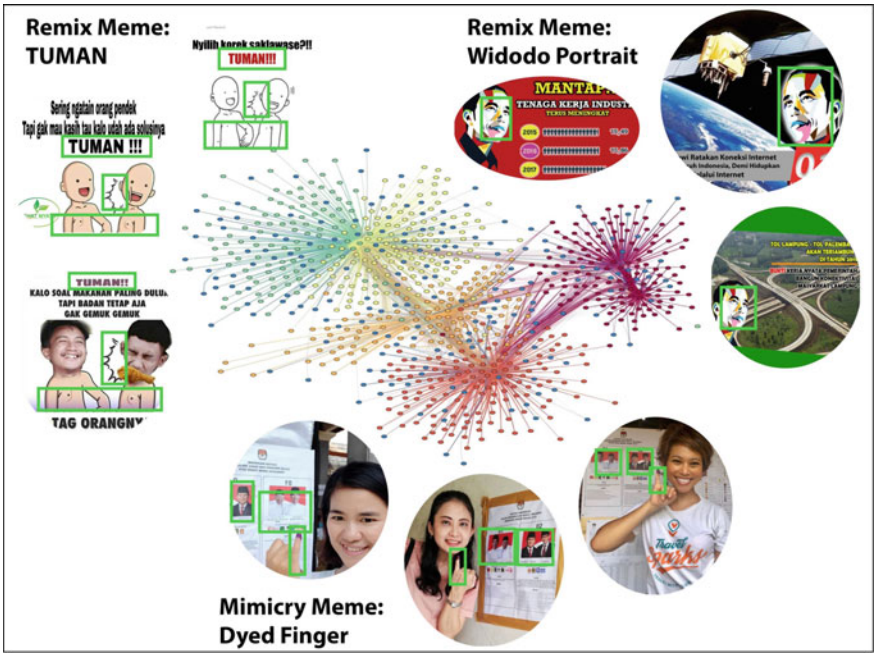


Fig. 15.20 A demonstration of the type of structure that may be inferred from the output of content retrieval, with many selected image queries. According to global and object-level local similarity, images with similar sub-components should be clustered together, presenting connections to their respective queries. Image adapted from William Theisen et al. (2020)

look into the dataset, it increases processing time for both the computer generating the clusters and the human reviewing them at the end. In their studies, Theisen et al. (2020) pointed out that 150 clusters appeared sufficient when the number of total dataset available images was on the order of millions of samples. This allowed for a maximum amount of images per cluster that a human could briefly review, but not so few images that trends could not be noticed.

15.4.2 Datasets and Evaluation

Indonesian Elections Dataset: The alternative content clustering stage inside the framework of provenance analysis aims to unveil emerging trends in large image datasets collected around a topic or an event of interest. To evaluate the capability of the proposed solutions, Theisen et al. (2020) collected over two million images from social media, concentrated around the 2019 Indonesian presidential elections. Harvested from Twitter (2021) and Instagram (2021) over 13 months (from March 31, 2018, to April 30, 2019), this dataset spans an extensive range of emotion and

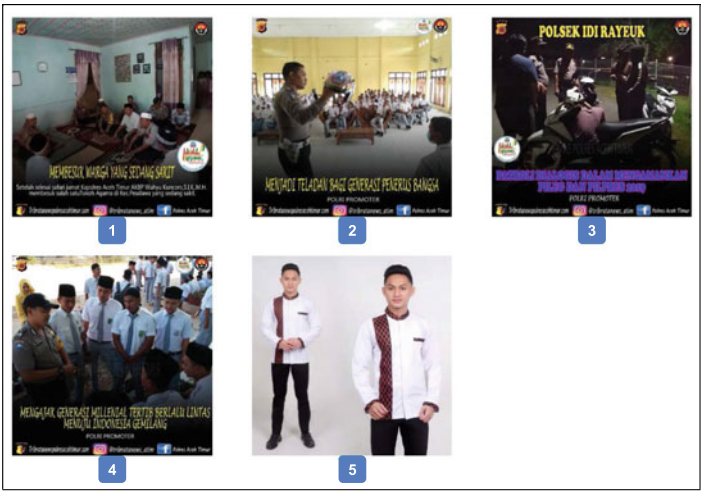


Fig. 15.21 Imposter-host test to measure the quality of the computed image clusters. Five images are presented at a time to a human subject, four of which belong to the same host cluster. The remaining one belongs to an unrelated cluster, therefore, being named imposter. The human subject is always asked to identify the imposter, which in this example is the last image. A high number of correct answers given by humans indicate that the clustering solution was successful. Image adapted from William Theisen et al. (2020)

thought surrounding the elections. The images are publicly available at https://bit.ly/2Rj0odI.

Human Understanding Assessments: Keeping the objective of aiding the human understanding of large image datasets in mind, a metric is needed to measure how meaningful an image cluster is for a human. Inspired by the work of Weninger et al. (2012), Theisen et al. (2020) proposed an imposter-host test, which is performed by showing a person a collection of N images, all but one of which are from a common "host" cluster, which was previously computed by the proposed solution, whose quality needs to be measured. The other item, the "imposter", is randomly selected from one of the other established clusters. The idea is that the more related the images in a single cluster are, the easier it should be for a human to pick out the imposter image. Figure 15.21 depicts an example of this configuration. To test the results of their studies, Theisen et al. (2020) hired Amazon Mechanical Turk workers (2021) to perform 25 of these imposter-host tasks each subject, with N equal to 5 images, in order to measure the ease with which a human can pick out an imposter from the clusters that were generated.

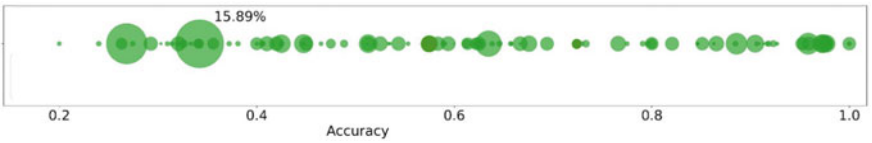


Fig. 15.22 Accuracy of the imposter-host tasks given to Amazon Mechanical Turk workers, relative to cluster size. The larger the point on the accuracy axis, the more images were in the cluster. The largest cluster is indicated as having 15.89% of all the dataset images. Image adapted from William Theisen et al. (2020)

15.4.3 Results

The Amazon Mechanical Turk experiments reported by Theisen et al. (2020) demonstrated that the provenance pipeline's clustering step can produce human interpretable clusters while minimizing the average cluster size. The average accuracy for the imposter-host test was 62.42%. If the worker is shown five images, the chance of correctly guessing would be 1/5 (20%). Therefore, the reported average is far above the baseline, thus, demonstrating the salience of the trends discovered in the clusters to human reviewers. The median cluster size was only 132 images per cluster. A spread of the cluster sizes as related to the worker accuracy for a cluster can be seen in Fig. 15.22. Surprisingly even the largest cluster, containing 15.89% of all the images, has an accuracy still higher than random chance. Three examples of what an individual cluster could look like can be seen in Fig. 15.23.

15.5 Open Issues and Research Directions

State of the Art: Based on the results presented and discussed in the previous sections within this chapter, one can safely conclude that, in the current state of the art of provenance analysis, the available solutions are indeed proven to be effective for at least two tasks, namely (1) authenticity verification of images and (2) the understanding of image-sharing trends online.

As for the verification of the authenticity of a questioned image (i.e. the query), this is done through the principled inspection of a given corpus of potentially related images to the query. In this case, whenever a non-empty set of related images is retrieved and a non-empty provenance graph (either directed or undirected) is generated and presented, one might infer the edit history of the query and take it as evidence of potential conflicts that might attest against its authenticity, such as inadvertent manipulations or source misattributions.

Regarding the understanding of trending content online, this can be done during the unfolding of a target event, such as national elections or international sports competitions. In this case, the quick retrieval of related images and subsequent content-based clustering have the potential to unveil emergent trends and surface

(a)

(b)

(c)

Fig. 15.23 Three examples of what the obtained clusters may look like. In **a** and **b**, it is very easy to see the similarity shared among all the images. In **c**, it is not quite as clear, but if one were to look a little more closely, they might notice the stylized "01" logo in each of the images. This smaller, shared component is what makes object-level matching and local feature analysis a compelling tool. Image adapted from William Theisen et al. (2020)

their provenance, allowing for a better comprehension of the event itself, as well as its relevance. A summary of what the public is thinking about an event and who the actors are trying to steer public opinion is only a couple of the possibilities that provenance content clustering may allow one to do.

While results have been very promising, no solution is without flaws, and much work is left to be done. In particular, image provenance analysis still has many caveats, which are briefly described below. Similarly, multimodal provenance analysis is an unexplored field, which deserves a great deal of attention from researchers shortly.

Image Provenance Analysis Caveats: There are open problems in image provenance analysis that still need attention from the scientific community. For instance, the solutions proposed to compute the directions of the edges of the provenance graphs still deliver values of edge overlap that are inferior to the results for vertex recall and vertex overlap. This aspect indicates that the proposed solutions are better at determining which images must be added to the provenance graphs as vertices, but there is still plenty of room to improve the computation of how these vertices must be connected. In this regard, novel techniques of learning which image might have been acquired or generated first and leveraging the output of single-image manipulation detectors (tackled throughout this book) are desired.

Moreover, an unexplored aspect of image provenance analysis is understanding, representing, and detecting the space of image transformations used to generate one image from the other. By doing so, one will determine what transformations might have been performed during the establishment of an edge, a problem that currently still lacks solutions. Thankfully, through the recent joint efforts of DARPA and NIST within the MediFor program (2021), a viable regime for data generation and annotation at the level of registering the precisely applied image transformations that have been performed from one asset to the other has emerged. This regime's outcome is available to the scientific community as a useful benchmark (see Guan et al. 2019) for further research.

Multimodal Provenance Analysis: As explained in Sect. 15.3, we have already witnessed that additional information such as image metadata helps to improve provenance analysis. Moving forward, another important research direction that requires progress is the identification and principled usage of information coming from other asset modalities (such as the text of image captions, in the case of questionable assets that are rich documents), as well as the development of provenance analysis for media types other than images (*e.g.*, the provenance of text documents, videos, audios, etc.).

When one considers the multimodal aspect, many questions appear and are open for research. One of them is how to analyze complex assets, such as the texts coming from different suspect documents (looking for cases of plagiarism and attribution to the original author), or the images extracted from scientific papers (which may be inadvertently reused to fabricate scientific findings, such as what was recently described in PubPeer Foundation (2020)). Another question is how to leverage images and captions in a document, or the video frames and their respective audio subtitles, within a movie.

Video provenance analysis, in particular, is a topic that deserves a great deal of attention. While one might assume image-based methods could be extended to video by being run over multiple frames, such a solution would fail to glean videos' inherent temporal dimension. Sometimes videos shared on social media such as Instagram (2021) or TikTok (2021) are composites of one or more pieces of viral footage. Tracking the provenance of videos is as important as tracking the provenance of still images.

Document provenance analysis, in turn, has also gained attention lately due to the recent *pandemic of bad science* (see Scheirer 2020). The COVID-19 crisis has caused a subsequent explosion in scientific publications surrounding the pandemic, not all of which might be considered highly rigorous. Using a provenance pipeline to aid in document analysis could allow reviewers to find repeated figures across many publications. It would be the reviewers' decision, though, to determine if the repetition is something as simple as citing the previous work or something more nefarious like trying to claim pre-existing work as one's own.

Towards an End-user Provenance Analysis System: Finally, for the solutions discussed in this chapter to be truly useful, they require front-end development and back-end integration, which are currently missing. Such a system would allow users to perform tasks similar to reverse image search, with the explicit intent of finding the origins of all related content within the query asset in question. If a front-end requiring minimal user skill to use could be designed, provenance analysis could be consolidated as a powerful tool for fighting fake news and misinformation.

Table 15.7 List of datasets useful for provenance analysis

Dataset	Year	Description	Link	QR Code
Reddit Photoshop Battles	2018	184 provenance graphs involving 10,421 images collected from the Reddit PhotoshopBattles community.	https://bit.ly/3jfrLSJ	
Open Media Forensics Challenge	2020	Reunion of NC17 MFC18, MFC19 and more recent sets, with useful partitions for provenance analysis.	https://bit.ly/3svh81R	
Indonesian Election Memes	2021	Over two million images collected from Twitter and Instagram about the 2019 Indonesian elections.	https://bit.ly/2Rj0odI	

15.6 Summary

The current state of the art of provenance analysis is mature enough for aiding image authenticity verification by processing a corpus of images rather than the processing of a single and isolated asset. However, more work needs to be done to improve the quality of the generated provenance graphs concerning the direction and identification of image transformations associated with the graphs' edges. Besides, provenance analysis still needs to be extended to media types other than images, such as video, audio, and text. In this regard, both the development of algorithms and the collection of datasets are yet to be done, revealing a unique research opportunity.

Provenance Analysis Datasets: Table 15.7 enlists the currently available and useful datasets for image provenance analysis.

Table 15.8 List of implementations of provenance analysis solutions

Solution	Year	Description	Link	QR Code
Context Incorporation	2017	Implementation of content retrieval based on context incorporation.	https://bit.ly/3daU2Ju	
Undirected Graph Construction	2017	Implementation of graph construction that generates undirected graphs.	https://bit.ly/3wRj2gM	
Provenance Analysis at Scale	2018	Implementation of content retrieval based on distributed keypoints and iterative filtering, and of graph construction that generates directed graphs, including the clustered approach.	https://bit.ly/3mKlVev	
NIST MedisScore	2018	Implementation of provenance analysis metrics and evaluation toolkit provided by NIST.	https://bit.ly/3anUMsN	
Leveraging Manipulation Detectors	2020	Implementation of content retrieval and of graph construction supported by image manipulation detectors.	https://bit.ly/3g44nZv	
Object-Level Content Retrieval	2021	Implementation of the OS2OS object-level content matching and retrieval strategy.	https://bit.ly/2Pv8OBM	
Transform-Aware Embeddings	2021	Implementation of graph construction based on transformation-aware embeddings.	https://bit.ly/3sgfrFo	

Provenance Analysis Source Code: Table 15.8 summarizes the currently available source code for image provenance analysis.

Acknowledgements Anderson Rocha thanks the financial support of the São Paulo Research Foundation (FAPESP, Grant #2017/12646-3).

References

Advance Publications, Inc (2012) Reddit PhotoshopBattles. https://bit.ly/3ty7pJr. Accessed on 19 Apr 2021

Amazon Mechanical Turk, Inc (2021) Amazon Mechanical Turk. https://www.mturk.com/. Accessed 11 Apr 2021

Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval, vol 463, Chap 8. ACM Press, New York, pp 191–227

Ballard D (1981) Generalizing the Hough transform to detect arbitrary shapes. Elsevier Pattern Recognit 13(2):111–122

Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Elsevier Comput Vis Image Understand 110(3):346–359

Bestagini P, Tagliasacchi M, Tubaro S (2016) Image phylogeny tree reconstruction based on region selection. In: IEEE international conference on acoustics, speech and signal processing, pp 2059–2063

Bharati A, Moreira D, Flynn P, Rocha A, Bowyer K, Scheirer W (2021) Transformation-aware embeddings for image provenance. IEEE Trans Inf Forensics Secur 16:2493–2507

Bharati A, Moreira D, Brogan J, Hale P, Bowyer K, Flynn P, Rocha A, Scheirer W (2019) Beyond pixels: Image provenance analysis leveraging metadata. In: IEEE winter conference on applications of computer vision, pp 1692–1702

Bharati A, Moreira D, Pinto A, Brogan J, Bowyer K, Flynn P, Scheirer W, Rocha A (2017) U-phylogeny: Undirected provenance graph construction in the wild. In: IEEE international conference on image processing, pp 1517–1521

Brogan J (2019) Reddit Photoshop Battles Image Provenance. https://bit.ly/3jfrLSJ. Accessed on 16 Apr 2021

Brogan J, Bestagini P, Bharati A, Pinto A, Moreira D, Bowyer K, Flynn P, Rocha A, Scheirer W (2017) Spotting the difference: Context retrieval and analysis for improved forgery detection and localization. In: IEEE international conference on image processing, pp 4078–4082

Brogan J, Bharati A, Moreira D, Rocha A, Bowyer K, Flynn P, Scheirer W (2021) Fast local spatial verification for feature-agnostic large-scale image retrieval. IEEE Trans Image Process 30:6892–6905

ByteDance Ltd (2021) About TikTok. https://www.tiktok.com/about?lang=en. Accessed 11 Apr 2021

Castelletto R, Milani S, Bestagini P (2020) Phylogenetic minimum spanning tree reconstruction using autoencoders. In: IEEE international conference on acoustics, speech and signal processing, pp 2817–2821

Chu Y-J (1965) On the shortest arborescence of a directed graph. Sci Sini 14:1396–1400

Costa F, Oikawa M, Dias Z, Goldenstein S, Rocha A (2014) Image phylogeny forests reconstruction. IEEE Trans Inf Forensics Secur 9(10):1533–1546

Costa F, De Oliveira A, Ferrara P, Dias Z, Goldenstein S, Rocha A (2017) New dissimilarity measures for image phylogeny reconstruction. Springer Pattern Anal Appl 20(4):1289–1305

Costa F, Lameri S, Bestagini P, Dias Z, Rocha A, Tagliasacchi M, Tubaro S (2015) Phylogeny reconstruction for misaligned and compressed video sequences. In: IEEE international conference on image processing, pp 301–305

Costa F, Lameri S, Bestagini P, Dias Z, Tubaro S, Rocha A (2016) Hash-based frame selection for video phylogeny. In: IEEE international workshop on information forensics and security, pp 1–6
De Oliveira A, Ferrara P, De Rosa A, Piva A, Barni M, Goldenstein S, Dias Z, Rocha A (2015) Multiple parenting phylogeny relationships in digital images. IEEE Trans Inf Forensics Secur 11(2):328–343
De Rosa A, Uccheddu F, Costanzo A, Piva A, Barni M (2010) Exploring image dependencies: a new challenge in image forensics. In: IS&T/SPIE electronic imaging, SPIE vol 7541, Media forensics and security II, pp 1–12
Dias Z, Rocha A, Goldenstein S (2012) Image phylogeny by minimal spanning trees. IEEE Trans Inf Forensics Secur 7(2):774–788
Dias Z, Goldenstein S, Rocha A (2013) Toward image phylogeny forests: automatically recovering semantically similar image relationships. Elsevier Forensic Sci Int 231(1–3):178–189
Dias Z, Goldenstein S, Rocha A (2013) Large-scale image phylogeny: tracing image ancestral relationships. IEEE Multimed 20(3):58–70
Dias Z, Goldenstein S, Rocha A (2013) Exploring heuristic and optimum branching algorithms for image phylogeny. Elsevier J Vis Commun Image Represent 24(7):1124–1134
Dias Z, Rocha A, Goldenstein S (2011) Video phylogeny: Recovering near-duplicate video relationships. In: IEEE international workshop on information forensics and security, pp 1–6
Edmonds J (1967) Optimum branchings. J Res Natl Bure Stand B 71(4):233–240
Facebook, Inc (2021) About Instagram. https://about.instagram.com/about-us. Accessed 11 Apr 2021
Ge T, He K, Ke Q, Sun J (2013) Optimized product quantization for approximate nearest neighbor search. In: IEEE conference on computer vision and pattern recognition, pp 2946–2953
Guan H, Kozak M, Robertson E, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: 2019 IEEE winter applications of computer vision workshops (WACVW). IEEE, pp 63–72
Idée, Inc. TinEye (2020) Reverse image search. https://tineye.com/. Accessed on 17 Jan 2021
Johnson J, Douze M, Jégou H (2019) Billion-scale similarity search with gpus. IEEE Trans Big Data 1–12
Kennedy L, Chang S-F (2008) Internet image archaeology: Automatically tracing the manipulation history of photographs on the web. In: ACM international conference on multimedia, pp 349–358
Kruskal J (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. Proc Am Math Soc 7(1):48–50
Lameri S, Bestagini P, Melloni A, Milani S, Rocha A, Tagliasacchi M, Tubaro S (2014) Who is my parent? reconstructing video sequences from partially matching shots. In: IEEE international conference on image processing, pp 5342–5346
Liu Y, Zhang D, Guojun L, Ma W-Y (2007) A survey of content-based image retrieval with high-level semantics. Elsevier Pattern Recognit 40(1):262–282
Lowe D (2004) Distinctive image features from scale-invariant keypoints. Springer Int J Comput Vis 60(2):91–110
Melloni A, Bestagini P, Milani S, Tagliasacchi M, Rocha A, Tubaro S (2014) Image phylogeny through dissimilarity metrics fusion. In: IEEE European workshop on visual information processing, pp 1–6
Milani S, Bestagini P, Tubaro S (2016) Phylogenetic analysis of near-duplicate and semantically-similar images using viewpoint localization. In *IEEE international workshop on information forensics and security*, pp 1–6
Milani S, Bestagini P, Tubaro S (2017) Video phylogeny tree reconstruction using aging measures. In: IEEE European signal processing conference, pp 2181–2185
Moreira D, Bharati A, Brogan J, Pinto A, Parowski M, Bowyer K, Flynn P, Rocha A, Scheirer W (2018) Image provenance analysis at scale. IEEE Trans Image Process 27(12):6109–6123
National Institute of Standards and Technology (2017) Nimble Challenge 2017 Evaluation. https://bit.ly/3e8GeOP. Accessed on 16 Apr 2021

National Institute of Standards and Technology (2018) Nimble Challenge 2018 Evaluation. https://bit.ly/3mY2sHA. Accessed on 16 Apr 2021
National Institute of Standards and Technology (2019) Media Forensics Challenge 2019. https://bit.ly/3susnrw. Accessed on 19 Apr 2021
Noh H, Araujo A, Sim J, Weyand T, Han B (2017) Large-scale image retrieval with attentive deep local features. In: IEEE international conference on computer vision, pp 3456–3465
Oikawa M, Dias Z, Rocha A, Goldenstein S (2015) Manifold learning and spectral clustering for image phylogeny forests. IEEE Trans Inf Forensics Secur 11(1):5–18
Oikawa M, Dias Z, Rocha A, Goldenstein S (2016) Distances in multimedia phylogeny. Int Trans Oper Res 23(5):921–946
Pinto A, Moreira D, Bharati A, Brogan J, Bowyer K, Flynn P, Scheirer W, Rocha A (2017) Provenance filtering for multimedia phylogeny. In: IEEE international conference on image processing, pp 1502–1506
Prim R (1957) Shortest connection networks and some generalizations. Bell Syst Tech J 36(6):1389–1401
PubPeer Foundation (2020) Traditional Chinese medicine for COVID-19 treatment. https://bit.ly/2Su3g8U. Accessed on 11 Apr 2021
Rudin C, Schapire R (2009) Margin-based ranking and an equivalence between adaboost and rankboost. J Mach Learn Res 10(10):2193–2232
Scheirer W (2020) A pandemic of bad science. Bull Atom Sci 76(4):175–184
Shifman L (2014) Memes in digital culture. MIT Press
Stella Y, Shi J (2003) Multiclass spectral clustering. In: IEEE international conference on computer vision, pp 313–319
Theisen W, Brogan J, Thomas PB, Moreira D, Phoa P, Weninger T, Scheirer W (2021) Automatic discovery of political meme genres with diverse appearances. In: AAAI International Conference on Web and Social Media, pp 714–726
Tian Y, Fan B, Wu F (2017) L2-net: Deep learning of discriminative patch descriptor in Euclidean space. In: IEEE conference on computer vision and pattern recognition, pp 661–669
Turek M (2020) Media Forensics (MediFor). https://www.darpa.mil/program/media-forensics. Accessed on 24 Feb 2021
Twitter, Inc (2021) About Twitter. https://about.twitter.com/. Accessed 11 Apr 2021
University of Notre Dame, Computer Vision Research Laboratory (2018) MediFor (Media Forensics). https://bit.ly/3txaZU7. Accessed on 16 Apr 2021
Verde S, Milani S, Bestagini P, Tubaro S (2017) Audio phylogenetic analysis using geometric transforms. In: IEEE workshop on information forensics and security, pp 1–6
Weninger T, Bisk Y, Han J (2012) Document-topic hierarchies from document graphs. In: ACM international conference on information and knowledge management, pp 635–644
Yi KM, Trulls E, Lepetit V, Fua P (2016) LIFT: learned invariant feature transform. In: Springer European conference on computer vision, pp 467–483
Zhang X, Sun ZH, Karaman S, Chang S-F (2020) Discovering image manipulation history by pairwise relation and forensics tools. IEEE J Sel Top Signal Process 1012–1023
Zhu N, Shen J (2019) Image phylogeny tree construction based on local inheritance relationship correction. Springer Multimed Tools Appl 78(5):6119–6138

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.