Multiple Parenting Phylogeny Relationships in Digital Images

Alberto A. de Oliveira, Pasquale Ferrara, Alessia De Rosa, *Member, IEEE*, Alessandro Piva, *Senior Member, IEEE*, Mauro Barni, *Fellow, IEEE*, Siome Goldenstein, *Senior Member, IEEE*, Zanoni Dias, and Anderson Rocha, *Member, IEEE*

***Abstract*— Recently, several studies have been concerned with modeling the parenthood relationships between near duplicates in a set of images. Two images share a parenthood relationship if one is obtained by applying transformations to the other. However, this is not the only form of parenting that can exist among images. An image might be a composition created through the combination of the semantic information existent in two or more source images, establishing a relationship between the sources and the composite. The problem of identifying these relations in a set containing near-duplicate subsets of source and composition images is referred to as multiple parenting phylogeny. Thus far, researchers tackled this problem with a three-step solution: 1) separation of near-duplicate groups; 2) classification of the relations between the groups; and 3) identification of the images used to create the original composition. In this work, we extend upon this framework by introducing key improvements, such as better identification of when two images share content, and improved ways to compare this content. In addition, we also introduce a new realistic professionally created data set of compositions involving multiple parenting relationships. The method we present in this paper is properly evaluated through quantitative metrics, established for assessing the accuracy in finding multiple parenting relationships. Finally, we discuss some particularities of the framework, such as the importance of an accurate reconstruction of phylogenies and the method's behavior when dealing with more complex compositions.**

***Index Terms*— Image forensics, multimedia phylogeny, multiple parenting phylogeny, compositions.**

Manuscript received April 8, 2015; revised August 14, 2015 and October 2, 2015; accepted October 6, 2015. Date of publication October 26, 2015; date of current version December 9, 2015. This work was supported in part by the Coordination for the Improvement of Higher Education Personnel-CAPES through the DeepEyes Project, in part by the National Council for Scientific and Technological Development-CNPq under Grants 304352/2012-8, 477662/2013-7, 306730/2012-0, 477692/2012-5, and 483370/2013-4, in part by São Paulo Research Foundation-FAPESP under Grants 2010/05647-4, 2011/22749-8, 2013/21251-1, and 2013/08293-7 and, by the European Union through the Reverse Engineering of Audio-Visual Content Data Project, which was supported by the Future and Emerging Technologies Programme within the Seventh Framework Programme for Research of the European Commission, under Grant FET-268478. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. H. Vicky Zhao.

A. A. de Oliveira, S. Goldenstein, Z. Dias, and A. Rocha are with the Institute of Computing, University of Campinas, Campinas 13083-852, Brazil (e-mail: alberto.a.oliv@gmail.com; siome@ic.unicamp.br; zanoni@ic.unicamp.br; anderson.rocha@ic.unicamp.br).

P. Ferrara, A. De Rosa, and A. Piva are with the Department of Information Engineering, University of Florence, Florence 50125, Italy (e-mail: pasquale.ferrara@unifi.it; alessia.derosa@unifi.it; alessandro.piva@unifi.it).

M. Barni is with the Department of Information Engineering and Mathematics, Universita di Siena, Siena 53100, Italy (e-mail: barni@dii.unisi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2015.2493989

I. Introduction

NOWADAYS, countless images, videos, texts, sounds, and all types of media are shared and reproduced, with a multitude of intentions. Once content is shared on the internet, copies, remixes, compositions and re-sharing are its most common fate, sometimes with a complete different purpose. Multimedia files sharing the same content, but diverging by slight transformations, receive the name of near duplicates. The problem of Near-Duplicate Detection and Recognition, referred to as NDDR, was vastly studied in the literature [1], [2], with particular interest in the image domain [3]–[5]. Yet, most of the existing NDDR systems have focused on simply discovering if two images are near duplicates, not exploring any causal relationships between them.

On the contrary, when an image is transformed generating a near duplicate, a parenting relationship exists between the two images, i.e., the former is parent of the latter. Some research groups started to take an interest in finding all the parenting relationships existent in a set of *Near-Duplicate Images* (NDIs). Kennedy and Chang [6] were the pioneers in this field of study by proposing the modelling of these relationships as a graph, which would be part of a larger, image archaeology system. However, their work lacked a suitable metric to measure how likely an NDI is parent of another. Dias et al. [7], [8] and De Rosa et al. [9] further improved upon this idea, by separately introducing quantitative ways of measuring parenting between NDIs. In Dias et al. work, the graph representing the causal relationship among the analyzed NDIs was a tree, named *Image Phylogeny Tree* (IPT).

Successively, the same authors extended their study considering a set of *Semantically Similar Images* (SSIs), a generalization of NDIs: SSIs are images with closely related content, but not necessarily obtained by applying a sequence of transformations to the same original image. For instance, consider two pictures taken from the same scene, but from different cameras, different moments in time, or from a slightly different point of view. These two sources, with their respective NDIs obtained by applying transformations to them, form an SSI set. Fig. 1 depicts a small SSI set example. When analyzing a set of SSIs, instead of a single Image Phylogeny Tree, Dias et al. aimed at reconstructing an *Image Phylogeny Forest* (IPF) [10], a set of IPTs, each rooted in one of the original images in the SSI set. Then Costa et al. [11] introduced several new approaches for the IPF problem.

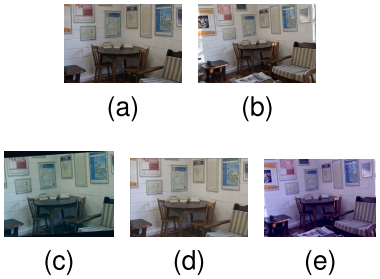

Fig. 1. Example of a set of Semantically Similar Images (SSIs) and Near-Duplicate Images (NDIs). (a) and (b), are original Images. (c) and (d) are NDIs obtained from (a) while (e) is an NDI obtained from (b).

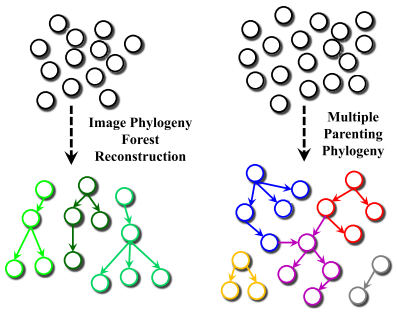

Fig. 2. An Image Phylogeny Forest reconstructs a forest to represent the entire SSI set. Multiple Parenting Phylogeny (MPP) reconstructs multiple trees, finding, if existent, a relationship between a composition and the images used to create it. Colors represent the relationships between images. In the forest scenario, the same color in different tones show how semantically similar images are in the same scene but with a different acquisition process. In the MPP scenario, the mixing of the red and blue colors result in the purple color, used to represent the composition process.

In all these works, the authors considered that images could only have one of its near duplicates as parent. However, there are cases when one image is created through the combination of parts from different images. This case is common, for example, in the form of splicings, montages, mosaics, and others, for which we use the general name of *compositions*. When two or more source images are used to create a new image, the composition has a relationship with the corresponding source images. This problem was addressed for the first time in [12] and dubbed *Multiple Parenting Phylogeny*, whereby the objective, given a set of images with varied contents, is to discover the different phylogenies and establish multiple parenting relationships in the set, if any. Fig. 2 illustrates the main objective of Multiple Parenting Phylogeny (MPP), versus Image Phylogeny Forest.

The increasingly frequent occurrence of image compositions on the internet renders the applications of Multiple Parenting Phylogeny very useful in practical scenarios such as content tracking, forensics and copyright enforcement. It is possible, for example, to enforce copyright by discovering whether an image was generated by using parts of protected images. More generally, provided we work on a framework capable of handling a large number of images, it is possible to track the way image contents transform on the web, and how they are used to create new ideas.

As already mentioned, the first work addressing MPP problem was presented in [12], whereby a 3-step approach laid out the foundation for addressing the MPP problem. However, without significant improvements in the various steps of the proposed approach, the results were just preliminary and the method not ready for dealing with more complex scenarios. This paper goes significantly beyond previous work, and this is demonstrated by testing the new version of the algorithm against several datasets, including photo realistic compositions. In the following, we clarify the novelties of this work by listing the key differences with respect to [12].

- For separating the different NDI sets, including parent and composition images, an exact optimum-branching method (i.e., Extended Automatic Optimum Branching, E-AOB) proposed by Costa et al. [11] for the Image Phylogeny Forest reconstruction has been explored. As we will see, although the E-AOB approach is shown to be the best in scenarios with Semantically Similar images, this is not always true for Multiple Parenting Phylogeny.
- To determine which IPT is the one including composition images, we combine the detection of a large number of local matches between images with a novel quality-based match selection approach, seeking to improve the correct detection of "reused material" shared between images. This new approach comes in handy when the composition is created by inserting a small region of an image (parent B) into another image (parent A), to find the match between the composition and the parent B.
- Regarding how to find the specific nodes (within the two parent IPTs) which generated the composition, we introduce a method for locally estimating the similarity between images that relies on the shared parts of such images, only. This method allows a better study of the relationships between two images when they have some parts in common, and it is indifferently applied to the two parent IPTs, so that no longer we need to recognize what kind of parent is (e.g., parent A or B in the previous examples), contrary to what happened in [12].
- Finally, in this work, we present a larger set of experiments to validate both the approach and the individual implementation steps of the overall Multiple Parenting Phylogeny system. We introduce a professionally-made Multiple Parenting Phylogeny dataset, comprising high resolution and more realistic compositions, to better evaluate the proposed method in a scenario closer to reality, also considering when some nodes of the different IPTs are missing. Furthermore, we test the method on real images coming from the web and discuss its potential and possible improvements.

The remaining of this paper has five sections. Section II discusses Image Phylogeny and its foundational concepts derived by previous literature. Section III introduces the proposed method for Multiple Parenting Phylogeny. Section IV details the datasets and experimental details. Finally, Section V presents the experiments and results while Section VI concludes the paper and discusses possible future research.

II. Image Phylogeny Concepts and Related Work

In this section, we provide the main concepts in the related literature that are the basis of the proposed work. We focus

on works in Image Phylogeny Trees, Image Phylogeny Forests and Multiple Parenting Phylogeny.

A. Image Phylogeny Tree

Image Phylogeny aims at reconstructing the existing relationships between images in a set of Near Duplicate Images, representing them as an Image Phylogeny Tree. In this vein, Dias et al. [7], [8] proposed a method comprising two main steps: the dissimilarity matrix estimation and the tree reconstruction.

The *dissimilarity* from an image I_A to an image I_B is an asymmetric measure used to estimate how likely it is for I_A to be the parent of I_B in the tree. Given the matrix containing all dissimilarity values, the reconstruction algorithm extracts its minimum directed spanning tree, which is the Image Phylogeny Tree.

1) Dissimilarity Matrix Estimation: Given two NDIs I_A and I_B, to compute their dissimilarity d_{I_A,I_B} it is important to define a family of image transformations $\mathcal{T}$, that could have been applied to I_A to obtain I_B. The first set of image transformations considered by Dias et al. [7] included scaling and JPEG compression and was later expanded [8] to also contemplate resampling, cropping, affine warping, color transformations (brightness, contrast and gamma correction) and compression. Let $T_{\vec{\beta}}$ be a subset of transformations belonging to $\mathcal{T}$, with parameter values $\vec{\beta}$. The dissimilarity d_{I_A,I_B} from image I_A to image I_B is defined by the following equation:

$$d_{I_A,I_B} = \min_{T_{\vec{\beta}}} \left| I_B - T_{\vec{\beta}}(I_A) \right|_{\text{pointwise comparison } \mathcal{L}}, \qquad (1)$$

considering all possible parameter values of $\vec{\beta}$. The minimization in Equation 1 requires $T_{\vec{\beta}}$ to be the transformation that best approximates I_A to I_B, according to the family of transformations $\mathcal{T}$, parameterized by $\vec{\beta}$. Thus, we estimate $T_{\vec{\beta}}$ and apply it to I_A, mapping it onto I_B's domain. Thereafter, we compute the dissimilarity as the difference between both images using any pointwise image similarity metric $\mathcal{L}$, such as mean squared error. The transformations estimated from I_A to I_B and from I_B to I_A are different, making it asymmetric.

To find $T_{\vec{\beta}}(I_A)$, Dias et al. [8] proposed the application of several transformation estimators, according to the family of transformations $\mathcal{T}$ described before, in the following order:

- **Resampling, cropping and affine warp:** geometric transformations that map I_A onto I_B's domain are estimated robustly using a Random Sample Consensus (RANSAC) [13]. The matching keypoints from both images used as input to the RANSAC are detected and described using Speeded-up Robust Features (SURF) [14];
- **Color transformations:** once the images are registered, the color transformations are estimated by normalizing each color channel of I_A using the mean and standard deviation of the corresponding color channel of I_B;
- **JPEG compression:** finally, image I_A is compressed using I_B's quantization table.

Once $T_{\vec{\beta}}(I_A)$ is found, *Minimum Squared Error* (MSE) is applied as point-wise comparison method to find the residual between the transformed I_A and I_B, which is the value of the dissimilarity d_{I_A,I_B}. This process is repeated for every pair of images contained in the analyzed set, in both directions, resulting in an $n \times n$ dissimilarity matrix M, where n is the number of NDIs in the set.

2) Phylogeny Tree Reconstruction: Let G be a complete, directed graph, with weights on its edges, induced by M, describing the parenting relationships among all pairs of NDIs in the set. The IPT is defined as the minimum spanning tree obtained from G. The first method proposed by Dias et al. [7], [8] to reconstruct the IPT was the *Oriented Kruskal*, a modification of Kruskal's minimum spanning tree classical algorithm [15] to operate on directed graphs. Oriented Kruskal is a greedy algorithm that starts considering each node of G as a possible root of the IPT. It then sorts the edges (dissimilarity values) of G, adding them to the IPT respecting the sorted order, avoiding cycles and loops, and joining trees if necessary. The IPT is complete when there is a single tree and every NDI in the set has been added to this tree.

Dias et al. later introduced [16] two new approaches for building the phylogeny tree. The first is an adaptation of Prim's algorithm [17] for finding the minimum spanning tree of a graph, which the authors named *Best Prim*. In general, the results did not outperform the ones obtained through Oriented Kruskal mentioned above. The other algorithm proposed is based on *Optimum Branching* (OB), proposed separately by Chu and Liu [18], Edmonds [19] and Bock [20]. As OB expects as input the root of the tree and this is unlikely to be known in real-world problems, the authors proposed to either use every node as root and select the tree with least cost or adding a dummy node with minimum cost to all other nodes of the tree and removing it later.

B. Image Phylogeny Forest

To represent the relationships between a set of Semantically Similar Images (instead of NDIs), Dias et al. [10] introduced the concept of Image Phylogeny Forest (instead of IPT) and proposed a new approach, extending upon their previous work. While the dissimilarity matrix estimation step remains the same, they introduced the *Automatic Oriented Kruskal* (AOK) as an IPF reconstruction algorithm. AOK establishes a threshold $\mathcal{C}$, defined as:

$$\mathcal{C} = k \times \sigma + M(I_X, I_Y) \qquad (2)$$

where k is a cutoff parameter, σ is the standard deviation of the dissimilarity values of the edges already added to the forest and $M(I_X, I_Y)$ is the dissimilarity value of the last edge added to the forest. The algorithm adds edges to the forest in the same way as detailed in Section II-A2, but respecting a new rule: the edge is only added if its dissimilarity value is less than $\mathcal{C}$. When all the nodes are added to the forest, the number of trees is automatically found.

Costa et al. [11] then introduced several new approaches for the IPF problem, such as the *Automatic Optimum Branching* (AOB), which applies the same cutoff limit used in AOK to OB. Using the initial solution returned by AOB, the authors observed that OB could be applied again to each individual

tree in the forest, resulting in optimal edge combinations for each individual tree, giving rise to a new algorithm which they called *Extended Automatic Optimum Branching* (E-AOB).

C. Multiple Parenting Phylogeny

When an image is not derived as the transformation of another image, but as the composition of two (or more) images, we step beyond the problem of Image Phylogeny (in the form of IPT or IPF) to the problem of Multiple Parenting Phylogeny (MPP). In this case, the objetive is to identify relationships between the composition image and the parent images used to create it, in sets containing NDIs.

In [12], authors considered a type of composition (a.k.a. splicing) whereby a part of an image (referred to as *alien*) is copied and pasted into another image (referred to as *host*). The studied test cases considered a simplification of the general Multiple Parenting Phylogeny scenario, comprising three groups of NDIs: one containing alien images, one containing host images and one containing compositions. To solve the problem, a 3-step framework was proposed in order to separate groups of near-duplicate images, to classify each tree and, finally, to identify host and alien source images.

The separation of NDI groups was done by applying an Image Phylogeny Forest (IPF) algorithm [10]. It was used to simultaneously separate groups of near duplicates and also to reconstruct their phylogenies as IPTs. Next, the shared content between all images from each IPT was examined; the composition images are the only ones that share content with both alien and host images, a fact that can be used to find out which one is the composition IPT. For finding shared content between images, it was used a keypoint match clustering approach, commonly seen in copy-move detection works [21]. Finally, after classifying the phylogeny trees as alien, host and composition trees, the root of the composition IPT (expected to be the original composition) was used as a reference for searching, in the other two IPTs, for the alien and host source images used to create it.

III. Proposed Method

In this section, we present a solution to deal with Multiple Parenting Phylogeny. We discuss the novelties introduced in the present approach by describing each phase composing the overall system. We also present a brief complexity analysis for all the steps involved in the method.

Within the Multiple Parenting Phylogeny scenario, we focus our interest in *splicing*-type compositions, created by combining the object of an *alien* image with the background of a *host* image. This is the most common type of composition observed in real scenarios, specially those of forensic nature. Fig. 3 depicts a splicing example.

In particular, we consider a set of images $\mathcal{S}$ containing three different groups of NDIs: host NDIs (S_H), alien NDIs (S_A), and composition NDIs (S_C).[1] The NDIs in each subset form an IPT for host, alien, and composition images, named T_H, T_A, and T_C, respectively. The root of T_C, named r_C is the original composition, which is created by combining an image $p_H \in T_H$ with an image $p_A \in T_A$. Fig. 4 shows an example of this scenario. In this vein, the MPP objectives are twofold: identifying T_H, T_A, and T_C, from the large set $\mathcal{S}$; and finding the images involved in the composition creation (r_C, p_H and p_A).

To achieve such objectives, we employ the 3-step framework, including: (1) forest reconstruction; (2) tree classification; and (3) parent identification; (c.f., Fig. 5). We now turn our attention to each step of the method.

[1]Note, however, that this is a simplification for explanation purposes. The framework is expected to work in the presence of unrelated images as well.

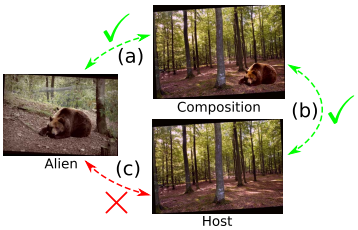

Fig. 3. Example of content relationships. (a) Alien and composition share the bear. (b) Host and composition share the entire background, minus the bear. (c) No content is shared (cut and pasted from one image to another) between alien and host.

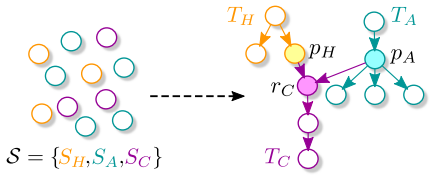

Fig. 4. A multiple parenting scenario for splicing compositions. The studied set of images comprises three subsets of host, alien and composition NDIs. The subsets form IPTs, whereby two random nodes of the host and alien IPTs where used to create the root of the composition IPT.

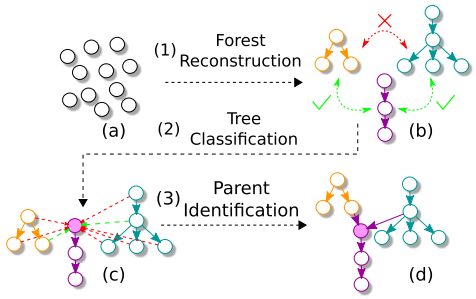

Fig. 5. Multiple Parenting Phylogeny framework. (a) Set of images. (b) After separating the groups of near-duplicate images, we search for content relationships between trees in order to classify them. (c) The purple colored tree in the center, having a content relationship with the other two, is the composition tree. (d) We search for the host and alien parent of the composition root in the remaining trees in order to establish the ancestors used to create the composition.

A. Forest Reconstruction

In this step, we aim at separating the NDIs existent in the set of analyzed images into groups of NDIs. This separation enables us to select specific images within each group to represent the whole group, as all NDIs share the same content. Moreover, we also restrict any future comparisons to non-NDIs, as we do not compare two images from the same group. Although there are multiple ways to identify

which images belong to the same NDI group, we apply an Image Phylogeny Forest algorithm. The main advantage of such an approach is to be able to simultaneously identify which images belong to the same NDI group and to reconstruct their phylogenies. The rationale for using an IPF algorithm is that since it works reasonably well when non-NDIs are very similar (as in the case of semantically similar images), it should also work when non-NDIs are completely different or only partially similar.

It is crucial to have an IPF approach capable of reconstructing with high accuracy the correct number of trees, with correct roots, and minimum mixing of nodes. The wrong reconstruction can lead to a strong negative effect in the subsequent steps of the method, specially in the identification of the original composition and its tree, and the classification of the host and alien trees.

In this work, in addition to *Automatic Oriented Kruskal* (AOK) [10], we also explore phylogeny forests reconstructed with the exact algorithm *Extended Automatic Optimum Branching (E-AOB)*. Introduced by Costa et al. [11], E-AOB is currently the state-of-the-art approach for IPF reconstruction (not involving combination of approaches). We present results comparing the reconstruction quality of AOK and E-AOB. We also discuss the results in an hypothetical scenario of a perfect forest reconstruction.

B. Tree Classification

With the groups separated as IPTs, the next step of the framework aims at identifying to which IPT belongs the composition, alien, and host images. Here, we employ a method based on analyzing the content relationships between candidate images of each IPT. Images I_A and I_B have a content relationship if part of the content of I_A is shared with I_B. Fig. 3 shows the content relationships between composition, host, and alien images. As the figure depicts, composition is the only image with a content relationship with the other two images. Thus, it is possible to classify the trees based on the content relationship observed between one of its members and members of the other trees.

To detect the aforementioned shared content between images, we apply a match clustering approach similar to the one used by Amerini et al. in their copy-move forgery detection work [21]. In that work, the authors present a technique based on J-Linkage [22] model estimation algorithm to cluster local matches between images, computed using *Scale Invariant Feature Transform* (SIFT) [23] features, based on their geometrical transformation. The principle is that matched keypoints inside a duplicated region will exhibit the same geometrical transformation, and hence can be clustered using this information. The same method can be applied to image parts used to create a composition. J-Linkage was chosen due to its robustness to false matches (important when looking for shared content between composition and alien images), as well as the usage of transformation properties, instead of easily deformable spatial properties, for matches clustering.

J-Linkage's accuracy and efficiency in detecting shared content between images is directly affected by the number of keypoints found as well as the quality of their matching. The SIFT extraction algorithm used for the task of keypoint detection and description is characterized by some parameters, such as peak and edge thresholds. By changing the value of such parameters, we can increase or decrease the number of detected keypoints. More deeply, by employing constraining thresholds, we reduce the number of detected keypoints, thus avoiding false matches and decreasing J-Linkage's execution time (which depends on the number of matches). Although less keypoints lead to a smaller number of false matches, it also decreases our chances of finding matches inside the shared content region of two related images. In this work, we choose an inverted approach: we relax such parameters in order to increase the chances of finding matches belonging to shared content between images.

Naturally, the increased number of keypoints generates a large number of false and redundant matches. When comparing compositions with aliens, the majority of matches are false, as in most cases both do not share much content. Hosts and compositions, on the other hand, have many redundant matches, unnecessary to finding shared content between them. In addition to relaxing the SIFT detection parameters, we introduce a quality-based *match selection* step, aiming at pruning the number of found matches. This step removes such false and redundant matches, improving J-Linkage's efficiency, while also keeping the necessary matches to detect shared content between matches. The efficiency improvement is particularly important, as detecting and extracting shared content between images is a recurrently used procedure in our framework.

The main advantage of adding a match selection step is that we can take away the burden of selecting which are the relevant matches from the keypoint detection, which has no information about match quality between images. Thus, we can keep only matches which we believe have higher chances (due to their quality) of being inside the shared content region. The quality measure chosen to be used was the distance between the matched points. First, we define a maximum number to be taken from the total amount of matches found in the SIFT keypoint matching procedure, by choosing only the c best matches (with smallest distances). Then, from the remaining matches, we define a cutting distance $\mathcal{D}$ as:

$$\mathcal{D} = \mathcal{M} + f \times \sigma \tag{3}$$

where $\mathcal{M}$ is the mean distance of the remaining matches, σ its standard deviation, and f a multiplicative factor. Any match whose distance is higher than $\mathcal{D}$ is excluded. Thus, the values of c and f control how rigid is the filtering process. The optimal values of both parameters are decided using a training set, a process which will be detailed in Section V-B.

By selecting the matches, we observed an increase in J-Linkage's efficiency, and, in many cases, in the accuracy of its shared content detection. This is most likely because of the reduction in false positives, contributing to a more precise detection of content relationships between images.

To find out which is the composition IPT, we use the aforementioned property that composition images share content with both alien images and host images. We pick random

images from each IPT to represent them, and analyze such images pairwise. If shared content is found between the pair, we increase a counter for the IPTs where the images originated from. We repeat this random pick procedures a total of five times, adding robustness to failures in detecting shared content between images. The choice of five repetitions here was enough for obtaining good results but more repetitions could be considered as well. The IPT with the highest count of content relationships found is classified as the composition IPT. Additionally, the root of the composition IPT is considered to be the original composition, created by combining one of the alien images with one of the host images.

With the composition tree classified, we use the dissimilarity value, computed to reconstruct the forest in Step 1, between the composition root and the roots of the remaining trees, in order to classify them. The dissimilarity is computed globally on the images, and thus, lower values indicate that the roots share much content (composition and host), while higher values indicate the reverse (composition and alien).

When there are more than three trees (host, alien and composition) present in the analysis such as when we have unrelated nodes or groups of NDIs broken into multiple trees, we consider for classification the IPTs with nodes that presented shared content relationships with nodes from the newly classified composition tree.

In scenarios with host and alien IPTs broken into multiple smaller IPTs, we cannot use the dissimilarity alone to classify the trees, as there will be multiple IPTs in need of classification. Instead, we can use an approach based on the covering of the shared content region between the roots (i.e., how much material they share). Consider that we just classified the composition tree and have multiple candidate trees to be classified into host or alien. We take the roots from the candidate trees and find the shared content between them and the composition tree. The ones whose shared content covers more than a threshold t of the composition root are classified as host, otherwise, they are classified as alien. If no shared content is found, we discard them. This approach can directly be used in a totally uncontrolled scenario such as the one containing images from the web, as we will show.

Although it is mostly true that a host and composition pair share a lot of content, while an alien and composition pair share only a small part, it is important to clarify that this is a simplification. Two exceptions are, for example, when a host undergoes severe cropping, leading to it having only a small part in common with a composition, and when the composition is cropped around the alien object, resulting in a lot of content in common with the alien image. However, given that we compare many different host and alien NDIs with the composition NDIs, the cropping bears little effect in the overall results of the method.

C. Parent Identification

Finding the correct host and alien nodes used to create the root of the composition tree is the last step of the multiple parenting framework. The approach for finding the host parent, although having yielded good results, depended heavily upon the dissimilarity value, computed in a different step of the method and in more practical scenarios might not always be available. Meanwhile, our initial results reported in [12] for identifying the alien parent identification results were very poor, even considering simple types of compositions. Moreover, the use of different methods for host and alien images was another issue, as the classification of trees is not always accurate, and might be even harder in more complex scenarios. We needed an approach that could be used with both host and alien images, while providing better parent finding results. To do this, we introduce a new method for computing the dissimilarity of two images, which considers only the content that is shared between them. The method, henceforth referred as Local Dissimilarity (LD), is divided in three parts:

1) Shared content identification;
2) Shared content region extraction;
3) Local Dissimilarity estimation.

Fig. 6. Registration and extraction of shared content region. (a) Original alien image. (b) Original composition. (c) Registered alien. (d) Shared content mask between the registered alien image and the composition image.

We start with the composition root r_C and a tree T (either host or alien), containing candidate images to be parent of r_C. Given a candidate image I_j from T, with $j \in [0, ..., |T| - 1]$, the identification of shared content between r_C and I_j proceeds in the same way as detailed in Section III-B: by using the J-Linkage match clustering approach [21], [22]. If shared content is not found between r_C and I_j, the LD of these images cannot be found, so we set it to infinite. Otherwise, we advance to the extraction step.

Extracting the shared content region of two images I_A and I_B consists in finding a binary mask R_j to represent the pixels in I_A that are shared with I_B. Only such pixels will be used in the estimation of the LD. In the previous step, we found a cluster of matches between r_C and I_j. By definition, true positive matched keypoints can only be found inside the shared content region in both images. Therefore, we can approximate the shared content region by using the convex hull of such matched keypoints in r_C. In addition, as the keypoints of r_C are used to estimate the mask, it is necessary to align the content of I_j to r_C, so the shared content region in both images overlap and can be compared using the mask R_j. Once again, we use the matches inside the cluster to estimate the transformation that aligns both images for later comparing the overlapped regions pointwise. The extraction procedure is exemplified in Fig. 6.

There is, however, an issue with the aforementioned approach. To find the parent, the values of the LD between r_C and all the candidate images in T will be compared, in search for the minimum. Yet, the binary masks found in the previous step often vary in size and shape for each image in T, thus

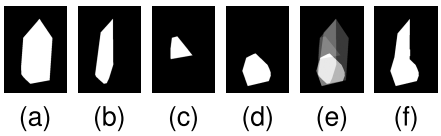

Fig. 7. Mask combination example. (a) to (d) Binary masks representing the shared content between alien parent candidate images and composition root. (e) Grayscale image representing the probabilities P_T of pixels belonging to the shared content region. (f) Final binary mask representing the shared content region between all alien images and the composition root.

making any measure computed inside them unsuitable to be compared, as they lack normalization. It is necessary to find a single binary mask $\mathcal{R}$ that represents the shared content region between the composition root r_C and *all* candidate images in T, as they are all NDI.

After finding all R_j, we introduce a statistical mask combination step to find $\mathcal{R}$, before computing the LD values. We start by computing a probability density function PDF_j to represent the probability of pixels in R_j of belonging to the shared content region between r_C and all images of T. PDF_j is estimated by using the frequentist definition of probability:

$$PDF_j(i) = \frac{R_j(i)}{\sum_i (R_j(i))} \tag{4}$$

where i is a pixel in mask R_j. After computing the PDFs for each mask, we combine probabilities of each mask R_j into a single probability function. To combine the masks we define a weight w_j for each mask, $w_j = \frac{N_j}{\sum_j (N_j)}$, where N_j is the number of matches found inside the cluster between I_j and the composition root r_C. The probability of each pixel belonging to the shared region is:

$$P_T = \sum_j (w_j \times PDF_j) \tag{5}$$

The binary mask $\mathcal{R}$ is obtained by taking the pixels with highest probabilities of P_T until the sum of their probabilities is equal or higher than a threshold p. Fig. 7 shows a simplified mask combination example with four masks, coming from alien parent candidate images.

After aligning all candidate images in T to r_C, and finding their combined mask $\mathcal{R}$, the last step is computing the LD. For image I_j from T, the LD between I_j and r_C is computed similarly to the dissimilarity measure discussed in Section II-A.1, by estimating the transformations that best approximate I_j to r_C, and using an image comparison measure. The main difference is that only the pixels inside $\mathcal{R}$ are used to estimate the transformations and to compute such measure. We use two improvements presented by Costa et al. [24] to compute the LD: histogram matching for color estimation and mutual information as the image comparison measure. Both showed the best results when compared to other approaches presented in [24]. However, both approaches are also considerably more costly than color channel normalization and MSE, thus we opt to not employ them in the global dissimilarity step, which requires several more image comparisons. The nodes p_H and p_A are the ones from the host and alien IPTs, respectively, with minimum LD to r_C.

D. Complexity Analysis

The forest reconstruction step comprises both the estimation of the dissimilarity matrix and the forest reconstruction. Given a set containing n images, estimating the dissimilarity matrix requires $n-1$ comparisons for each of the n images, resulting in $O(n^2)$ comparisons. According to Dias et al. [10], AOK has a complexity of $O(n^2 \log n)$ when using a combination of *disjoint-set-forest* with *union-by-rank* and *path-compression heuristics*. E-AOB's complexity is based on the original Optimum Branching algorithm [18]–[20]. Costa et al. [11] use an $O(n^2)$ implementation of OB. E-AOB runs OB again for each of the t trees found. Running OB again for each tree also has an upper bound of $O(n^2)$, as t is never greater than n, and neither is the number of images in each tree. Thus, the complexity of E-AOB is also $O(n^2)$.

For forest classification, we select one random node from each tree, and compare each pairing using J-Linkage totalling $\binom{t}{2} = \frac{t^2-t}{2}$ pairings with an an upper bound of $O(t^2)$ comparisons. As the number of trees is never greater than n, we have an upper bound of $O(n^2)$ on the number of images.

Finally, to identify the parents, we have to compare each of the alien and host nodes with the composition root. As the number of nodes in each tree will never be greater than n, the upper bound on the number of comparisons is $O(n)$. The three steps combined have upper bound of $O(n^2 + n^2 + n)$, and, therefore, the complete method has complexity of $O(n^2)$.

The bottleneck of the framework is related to J-Linkage. Besides being dependent on the input number of matches, it may also vary in how long it takes to converge. A small experiment was conducted to measure the average time taken by J-Linkage to search for a cluster of matches when comparing hosts to compositions and aliens to compositions. The results of such experiment are discussed in Section V.

IV. Experimental Setup

We now present the experimental setup and the test cases used for evaluating the methods. All datasets and test cases discussed in this section, as well as source code for testing and validation will be available for free through FigShare [25] (Datasets and test cases) and GitHub [26] (Source code).

A. Datasets

To evaluate the proposed multiple parenting approach, it is necessary to have available compositions as well as the source images used to create them. The two datasets comprising *splicing*-type compositions are discussed next.

1) Automatic Splicing Dataset: The first dataset used comprises splicing compositions created using two different pasting processes: *Direct Pasting* and *Poisson Blending*. They differ in the transformations applied to the object cut from the alien image and pasted onto the host's background to create the composition. Direct Pasting compositions have the object pasted onto the background with no further transformations applied. In Poisson Blending compositions, the alien's object is pasted onto the host using Pérez et al. [27] method of gradient matching, to better blend it to the host's background.

TABLE I
TRANSFORMATION PARAMETERS USED TO GENERATE NDIS

	Transformation	Operational Range
Geometric	Rotation	[-2.5°, 2.5°]
	Scale (Global or Per Axis)	[95%, 105%]
	Off-diagonal correction	[0.95, 1.05]
Color	Brightness adjustment	[-5%, 5%]
	Contrast adjustment	[-5%, 5%]
	Gamma adjustment	[0.95, 1.0]
	Cropping	[0%, 2.5%]
	JPEG Compression	[70%, 100%]

Each image is used as root for creating multiple different phylogeny trees. To generate the tree, we randomized a topology and sequentially transformed the root image to generate its nodes. *ImageMagick* [28] algorithms were used to generate NDI. Table I displays the transformations and their operational ranges used. Such transformations are cumulative: their values are randomized to generate the child of a node, and then randomized again to generate the child of the child. We followed Dias et al. [8] protocol for NDI generation, with slightly different operational ranges for transformations.

The dataset contains 100 host images of indoor and outdoor backgrounds, such as open fields, rooms or beaches. The images were obtained from the *Inria Holidays* [29] dataset. Also, the dataset contains 150 alien images of people, animals or cars, in common backgrounds. As the compositions were generated automatically, each alien image also accompanies a segmentation mask, describing the region to be cut. In most cases, the pasted alien object covers from 10% to 30% of the host image, with a few exceptions. The alien images were obtained from the *Berkeley Segmentation* [30] and *Graz-02* [31] datasets, and their segmentation masks from *Interactive Segmentation Tool* [32] and *Inria Annotations* [33], respectively. Each host and alien images are root to 25 different Phylogeny trees, varying within five tree topologies, each with five different sets of parameter values for generating the NDI in the tree. Each tree has 25 NDIs. The resulting compositions, host, and alien images are around 1MP.

In total, 5,000 compositions were created, equally divided into Direct Pasting and Poisson Blending. To create the compositions, after picking a pair of host and alien images, two Phylogeny Trees, one associated with each image, are randomly selected. After generating the trees, we randomly pick one host and alien images from each tree, using both to create the composition. All compositions were automatically created, and therefore the pasted object usually do not respect the semantic context of the background onto where it is pasted. Although 5,000 compositions were created, for evaluation purposes, we randomly selected a subset of them (for details, please refer to Section IV-B).

2) Professional Splicing Dataset: Although the previous dataset is a good starting point for studying splicing compositions, it has some limitations. The first, is the lack of a middle ground when considering the transformations the object undergo when pasted onto the background. With Direct Pasting, there are no transformations, resulting in the color and lighting of the object clashing with the background. In Poisson Blending, on the other hand, although the color of the object is somewhat adjusted to the background, in most cases the object ends up with a washed out transparency in relation to the background. In addition, the absence of context resulted from the automated composition creation process, makes both composition types not realistic enough for a human viewer.

With these problems in mind, a new dataset is proposed in this work, to provide compositions that look more realistic to human viewers. A new set of higher quality host and alien images was used, obtained from photo sharing websites such as *Flickr* [34] or media repositories such as *Wikimedia Commons* [35]. The host and alien images have around 2MP in size. A professional artist was hired to manually create each composition. The new set contains 100 host images, 27 of them belonging to the Inria Holidays [33] dataset and 73 collected from the internet. In addition, it has 64 aliens, all from the internet. As in the Automatic Splicing Dataset, the alien and host images all have 25 phylogeny trees, of 25 nodes each, associated with them. NDI were generated using the same procedure as the Automatic Splicing Dataset and the parameters of Table I.

This dataset has a total of 106 compositions, created in such a way that each host and alien was used at least once. The host and alien images were paired considering factors such as context, lighting and object angle in relation to the background. Once a pair of host and alien was chosen, a random tree associated to each is selected and their nodes generated. The NDIs comprising each tree were then provided to the artist, so he could select any two images from each tree that he found suitable to create the composition. The artist was free to transform the color, lighting and size of the pasted object, in order to create the most realistic compositions, provided the source images. The resulting compositions were further divided in three groups (easy, medium and hard). This classification was done according to how difficult it was to human viewer to perceive them as compositions, based on properties such as size of the object (both in the alien and in the composition images), color, illumination, and how well it blends the object onto the background. The sizes of the pasted object vary usually in the range of 5% to 30% of the host image total size, with a few exceptions depending on the background.

B. Test Cases

1) Baseline Test Cases: a total of three sets of test cases, using the base images and trees of the datasets formerly described were used, referred by the type of composition in it: Direct Pasting (DIP), Poisson Blending (PSB) and Professionally Made (PRM). Each test case is an Image Phylogeny Forest containing three Phylogeny Trees, the host IPT, the alien IPT and the composition IPT. As mentioned in the previous section, the composition is created by combining the content of two nodes from the host and alien IPTs. This composition is then used as root for the composition IPT, containing only its near duplicates. Each IPT has 25 nodes, adding to a total of 75 images per IPF in each test case. Fig. 4 depicts a small test case example, with fewer nodes in each IPT.

The DIP and PSB sets contain each a total of 400 cases, divided in 100 training and 300 testing examples. Training cases are used for calibration tasks, such as finding the best forest reconstruction parameter. For the PRM set, all the professionally made compositions are used, totaling 106 cases. Those cases are divided in 26 training and 80 testing cases. The cases belonging to training and testing were randomly selected, but in a way that the number of easy, medium and hard professionally made composition examples was balanced.

The aforementioned test cases are a slight simplification of multiple parenting phylogeny as they do not contain images unrelated to compositions, aliens, and hosts, and neither compositions created by combining more than two images. The next setups comprise more challenging scenarios with missing nodes and internet images.

2) Extended PRM Set: in addition to the aforementioned sets, we also have two additional extensions for the professionally made test case set. The extensions contemplate two scenarios: unrelated nodes and unrelated nodes plus missing nodes. In the former, we take the test cases of the PRM set containing 75 nodes of alien, host and composition images, and add 25 new unrelated images, which are themselves organized into 2 trough 5 new IPTs. This set is referred to as *PRM-EXP*, and each of its test cases have 100 images. The images added to the PRM-EXP set were selected from a pool of images from the *Inria Holidays* [29] dataset which were not used as host images.

For the missing nodes scenario, we take the PRM-EXP test cases and consider three variations, *CRin*, *HPin* and *APin*. In the *CRin* variation, we force keeping the composition root (CR) while forcing the removal of the host parent (HP) and alien parent (AP). In addition, we remove other 23 random nodes of the three. For the *HPin* and *APin* variations, we proceed similarly forcefully keeping the host parent and eliminating CR and AP (*HPin* case) and the keeping the alien parent and eliminating HP and CR (*APin* case). With the 23 randomly chosen nodes plus the two forced removals, each test cases retain, each one, 75 nodes.

3) Web Test Case: In addition to the controlled scenarios, we tested our approach in one test case containing images obtained from the web. This test case has a total of 56 images, with 26 hosts, 25 aliens and six compositions. NDIs were obtained using *TinEye* [36] and *Google Image Search* [37].

C. Evaluation

The evaluation of the proposed method is divided in two stages: forest reconstruction and multiple parenting. The first is concerned with the correct reconstruction of phylogenies in the set, according to Dias et al. metrics [10], as well as the separation of groups, measured through a metric designed specifically for this work. Multiple parenting evaluation, in turn, analyzes if the method can correctly identify which image is the original composition, as well as the correct host and alien images used to create it. Each evaluation stage is detailed next.

1) Forest Reconstruction Evaluation: this stage checks the quality of the IPF reconstructed in the first step using the forest reconstruction algorithm. The quantitative metrics (*roots*, *edges*, *leaves* and *ancestry*) used for such evaluation were defined by Dias et al. [10] and assume the *groundtruth* of the forest topology is available. The reconstructed and groundtruth IPF are denoted, respectively, as IPF^R and IPF^{GT}. The $Roots$ metric checks if the roots in IPF^R are the same as in IPF^{GT}, penalizing any additional or different roots found in IPF^R. The $Edges$ and $Leaves$ metrics measure the intersection of the edges and leaves set in IPF^R and IPF^{GT}. The $Ancestry$ metric checks if each node from IPF^R has the same set of ancestors as in IPF^{GT}. Formally, the aforementioned metrics are all defined as:

$$Metric(S^R, S^{GT}) = \frac{|S^R \cap S^{GT}|}{|S^R \cup S^{GT}|} \tag{6}$$

where $Metric$ can be any of the four metrics defined before ($Roots$, $Edges$, $Leaves$ and $Ancestry$), and S^R and S^{GT} is the set of the IPF elements related to such metrics in the reconstructed and groundtruth forests, respectively. For example, if $Metric$ is $Roots$, then S^R are the roots of IPF^R and S^{GT} the roots of IPF^{GT}. Moreover, if $Metric$ is $Leaves$, then S^R are the leaves of IPF^R and S^{GT} the leaves of IPF^{GT}.

A good separation of groups is, for multiple parenting, one of the most important features of the forest reconstruction step. It means that the host, alien, and composition near duplicates end up grouped together in the reconstructed forest, minimizing errors in subsequent steps. To evaluate this property, we propose the *subset* metric. It counts the proportion of nodes in IPF^R whose parent belongs to the same group (alien, host, or composition), excluding the root. For a more formal description, we start by defining three auxiliary functions. Let I_A be any image from the analyzed set, with parent I_B in the reconstructed forest IPF^R. $\pi(I_A, IPF^X)$ is a function that returns the parent of I_A in IPF^X. $\tau(I_A, IPF^X)$ is a function that returns the tree to which I_A belongs in IPF^X. Lastly, $\rho(IPF^X)$ is a function that returns the roots of IPF^X. We define $\sigma(IPF^R, IPF^{GT})$, the set of nodes from IPF^R with a parent which belongs to the same tree in IPF^{GT}, as:

$$\begin{aligned}\sigma(IPF^R, IPF^{GT}) = \{(I_A, I_B) | \pi(I_A, IPF^R) = I_B \\ \wedge \tau(I_A, IPF^{GT}) = \tau(I_B, IPF^{GT}), \\ \forall I_A \in IPF^R \setminus \rho(IPF^R)\}\end{aligned} \tag{7}$$

The *subset* metric is then defined as:

$$subset(IPF^R, IPF^{GT}) = \frac{|\sigma(IPF^R, IPF^{GT})|}{|IPF^R \setminus \rho(IPF^R)|} \tag{8}$$

2) Multiple Parenting Evaluation: To evaluate the multiple parenting reconstruction, we analyze three particular images of each test case (shown in Fig. 4): the composition root (r_C), the host parent (p_H) and the alien parent (p_A). These are the images involved in the composition creation process, which our method needs to identify. Three metrics, named after those images, are used to evaluate the multiple parenting reconstruction: Composition Root (CR), Host Parent (HP), and Alien Parent (AP). As the name suggests, the metrics check if the respective images found in the result are correct. In addition, we have three metrics named Composition Node (CN), Host Node (HN), and Alien Node (AN) to measure if the

classifications of the trees are correct. These metrics check if the composition root, host parent and alien parent found belong to the correct trees (composition tree, host tree and alien tree respectively). We now formally define all metrics.

Let the result of a given method be a triple (r_C^R, p_H^R, p_A^R) containing the composition root, host parent and alien parent found. Moreover, we also have available for evaluation purposes a triple $(r_C^{GT}, p_H^{GT}, p_A^{GT})$ containing the groundtruth composition root, host parent and alien parent. As defined in the previous section, $\tau(I_A, IPF^X)$ returns the tree to which image I_A belongs in IPF^X. The six aforementioned metrics are then defined as:

$$CR(r_C^R, r_C^{GT}) = \begin{cases} 1, & \text{If } r_C^R = r_C^{GT}. \\ 0, & \text{Otherwise.} \end{cases} \tag{9}$$

$$HP(p_H^R, p_H^{GT}) = \begin{cases} 1, & \text{If } p_H^R = p_H^{GT}. \\ 0, & \text{Otherwise.} \end{cases} \tag{10}$$

$$AP(p_A^R, p_A^{GT}) = \begin{cases} 1, & \text{If } p_A^R = p_A^{GT}. \\ 0, & \text{Otherwise.} \end{cases} \tag{11}$$

$$CN(r_C^R, r_C^{GT}) = \begin{cases} 1, & \text{If } \tau(r_C^R, IPF^R) = \tau(r_C^{GT}, IPF^{GT}). \\ 0, & \text{Otherwise.} \end{cases} \tag{12}$$

$$HN(p_H^R, p_H^{GT}) = \begin{cases} 1, & \text{If } \tau(p_H^R, IPF^R) = \tau(p_H^{GT}, IPF^{GT}). \\ 0, & \text{Otherwise.} \end{cases} \tag{13}$$

$$AN(p_A^R, p_A^{GT}) = \begin{cases} 1, & \text{If } \tau(p_A^R, IPF^R) = \tau(p_A^{GT}, IPF^{GT}). \\ 0, & \text{Otherwise.} \end{cases} \tag{14}$$

3) Multiple Parenting Evaluation With Missing Nodes: the scenarios of missing nodes detailed in Section IV-B2 requires a slightly different method of evaluation than the regular and expanded test cases. Because of missing nodes, a single IPT might be divided into multiple IPTs, thus any metric that searches for a correct root, such as the *CR* metric defined in Section IV-C2, should look into the set of new roots that appear. When the correct host or alien parents are removed, it does not make sense to look for it in the remaining nodes. Instead, we want to measure how far from the closest ancestor of the removed Host or Alien parent the algorithm has guessed. For instance, if a node is removed, it makes sense the algorithm to guess as the closest possible answer a immediate ancestor of such removed node that is still present in the tree.

The Composition Root (CR) metric becomes the Composition Root Set (CRS) metric, and its value is true if the r_C^R pointed by our algorithm belongs to the set of composition images that are roots in the IPF with missing nodes. Formally, it is defined as:

$$CR(r_C^R, r_C^{GT}) = \begin{cases} 1, & \text{If } r_C^R \in \mathcal{R}_C^*. \\ 0, & \text{Otherwise.} \end{cases} \tag{15}$$

where $\mathcal{R}_C^*$ is the set of composition images that became roots in the IPF with missing nodes.

For the Host Parent (HP) and Alien Parent (AP) metrics, we define the Host Parent Distance (HPD) and Alien Parent Distance (APD) metrics. First, we define the *Host Ancestor* and *Alien Ancestor* as α_H^{GT} and α_A^{GT}, respectively. Those are the closest surviving ancestors to the p_H^{GT} and p_A^{GT} in the tree with missing nodes. If p_H^{GT} was not removed in the tree with missing nodes, then $\alpha_H^{GT} = p_H^{GT}$ (the same is valid for the alien parent). If the entire branch, to the root, of p_H^{GT} was removed, then $\alpha_H^{GT} = \emptyset$ (similarly for the alien parent). The metric HPD is defined as:

$$HPD(p_H^R, p_H^{GT}) = edges^{GT}(p_H^R, \alpha_H^{GT}) \tag{16}$$

where the function $edges^{GT}(a, b)$ returns the number of edges between nodes a and b in IPF^{GT}. Metric APD is defined in a similar fashion. Fig. 8 illustrates this metric. However, to measure HPD, two conditions must be met: (i) the node found as host parent p_H^R must be a host node, which is exactly what the the HN metric checks, and (ii) $\alpha_H^{GT} \neq \emptyset$, meaning that at least one of the ancestors of p_H^{GT} was conserved in the tree with missing nodes. Both conditions also apply to APD.

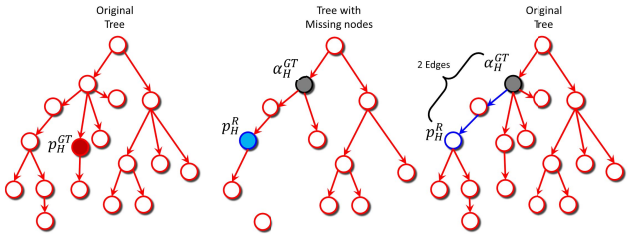

Fig. 8. Example of measuring the HPD metric. In this case, HPD = 2. The original tree is used for measuring the distance, as in the tree with missing nodes there might not be a path between p_H^R and α_H^{GT}.

V. Experiments and Results

This section shows the experimental results for the discussed methods. Recall that the baseline test cases are named after the type of composition contained in it: Direct Pasting (DIP), Poisson Blending (PSB) and Professionally Made (PRM). We also discuss the results for the extended PRM set, comprising the *expanded* scenario, in which several nodes of unrelated images are added to the test cases of the PRM set (PRM-EXP) and the missing nodes scenario, in which we take the test cases of the expanded set and further remove various random nodes (CRin, APin, HPin). Finally, we close this section with a real-world qualitative analysis of a host, alien, and composition scenario obtained from the web.

A. Forest Reconstruction Results

This section compares the forest reconstruction results of two different algorithms: Automatic Oriented Kruskal [10] and Extended Automatic Optimum Branching [11]. Both algorithms have an input parameter k related to the tree cutoff, that is, how likely the algorithms are able to separate a tree in two, which was detailed in Section II-B. The value of k is found for each test case set and forest building algorithm combination. To determine the value of k, we used the *training* portion of the sets. The algorithms were run with k values ranging from 0.0 to 10.0, in intervals of 0.2. The criteria used to select the

TABLE II

CHOSEN PARAMETERS IN TRAINING FOR FOREST RECONSTRUCTION AND MATCH SELECTION

	Forest Reconstruction		Match Selection	
	k		c	f
Set	AOK	E-AOB		
DIP	3.0	3.0	250	2.0
PSB	3.0	2.8	250	2.0
PRM	3.0	3.2	500	0.5

TABLE III

COMPARISON BETWEEN AOK AND E-AOB FOREST RECONSTRUCTION RESULTS FOR EACH COMPOSITION TYPE CONSIDERED HEREIN

		Metrics				
Set	Algorithm	Roots	Edges	Leaves	Ancestry	Subset
DIP	AOK	81.6%	74.3%	81.4%	65.5%	99.9%
	E-AOB	83.6%	76.1%	83.4%	68.0%	99.9%
Wilcoxon		✓	✓	✓	✓	×
PSB	AOK	78.7%	74.5%	81.3%	63.2%	99.7%
	E-AOB	80.3%	76.2%	83.5%	66.1%	99.8%
Wilcoxon		×	✓	✓	✓	✓
PRM	AOK	79.2%	74.9%	81.4%	61.9%	99.5%
	E-AOB	80.1%	77.8%	84.2%	65.5%	99.6%
Wilcoxon		×	✓	✓	✓	×

best parameter was its results in identifying the correct number of trees, followed by *roots* metric's performance, seeking for a balance of both. Table II shows the values of k used. Once k was determined, we reconstructed the forests from the *testing* portion of each test case set. Table III shows the results. The Wilcoxon signed-rank test was used to test for statistical significance at 95% confidence level. Green check marks point out results that are statistically significant, while red X's point out otherwise.

As Table III shows, for all types of compositions, the forest reconstruction results are good. For multiple parenting, two metrics are specially important: *roots* and *subset*. For all composition types, the results for *roots* were close to 80%, with E-AOB in all cases. This is specially important, as after identifying the phylogeny tree of the composition NDI, we choose its root as the original composition. However, note that the gains obtained by E-AOB in the roots metric were not statistically significant for PSB and PRM. The results for the subset metric were also interesting: close to 100% for all cases. This means that the method achieves, for all types of compositions, close to perfect separation of groups with both AOK and E-AOB.

B. Multiple Parenting Results

In this section, we discuss the multiple parenting results, measured by the metrics introduced in Section IV-C2. We start by comparing the results of the multiple parenting approach presented herein, using two techniques for separating the near-duplicate groups. Then, we present the results considering two setups: one simulating perfect reconstruction of forests, by using the groundtruth forests (GT) available, and another using forests reconstructed with E-AOB. The importance of this experiment is to highlight how errors in the forest reconstruction affects Steps (2) and (3) of the multiple parenting framework. Following, we show the results of the extended approach for the Professionally-Made Dataset, divided per difficult level of composition. This section is closed with a small experiment to assess J-Linkage's efficiency, when comparing aliens and compositions, as well as hosts and compositions.

The procedure for finding the match selection parameters c and f was similar to IPF reconstruction algorithm's parameter k. Using only the groundtruth forests and the *training* portion of each test case set, we grid searched the combinations of $c \in \{250, 500, 750\}$ and $f \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$, choosing the one yielding the best results while having reasonable efficiency, for each test set. Table II shows the best values of c and f for each test set.

Table IV shows the results of the multiple parenting approach presented in this paper using the two aforementioned forest reconstruction techniques to separate groups of near-duplicates. As all metrics are binary in this case, we used McNemar's test to check for statistical significance at 95% confidence level. For both reconstruction approaches, the proposed multiple parenting approach achieves good results. The method using E-AOB to separate the groups of NDIs slightly outperforms (with no statistical significance) the version using AOK. The good results for the tree classification metrics (CN, HN and AN) are mostly due to the shared content recognition step, which employs better keypoint localization and SIFT matching, as well as the match selection described in Section III-B. The very good CN, HN and AN results (over 85% in all cases), highlight the accuracy of the method in classifying the different NDI groups.

A better classification of trees directly improves the identification of the composition root. In our pipeline, the biggest challenge is identifying the Alien Parent (AP metric), as it involves localizing an often small patch of pasted content in a composition, and comparing this patch with the images we suspected it originated from. In particular, for the PSB and PRM test sets, this task is even harder, as the pasted content blends into the background. Yet, the method achieves over 50% accuracy in identifying this image in the PSB and PRM sets; a good result, considering that the identification of the Alien Image is related to a good separation of the trees and subsequent correct classification of the alien tree.

Table V compares the multiple parenting results in a fictitious informed scenario where we know the correct phylogenies in the set (named GT), and a scenario where such phylogenies are obtained by using E-AOB to reconstruct them. As in the previous Multiple Parenting results, McNemar's test was used to check for statistical significance. The use of the groundtruth here aims at simulating a *perfect* group separation thus isolating the impact of this factor in the overall analysis. In doing so, we have knowledge about the NDIs that are grouped together, but we do not know the type of each group, neither do we know the correct images involved in the multiple parenting. In this case, the proposed method is still responsible for classifying the separated groups, and identifying the parents within the groups.

The results displayed are important to isolate the errors arising from an incorrect reconstruction of the IPF, from

TABLE IV

RESULTS FOR OUR MULTIPLE PARENTING APPROACH BY COMPARING THE TWO FOREST RECONSTRUCTION METHODS, AOK AND E-AOB, FOR THE THREE COMPOSITION TYPES

	DIP						PSB						PRM					
Work	**CR**	**CN**	**HP**	**HN**	**AP**	**AN**	**CR**	**CN**	**HP**	**HN**	**AP**	**AN**	**CR**	**CN**	**HP**	**HN**	**AP**	**AN**
MPP E-AOB	81.00%	99.30%	91.00%	99.30%	86.00%	100.00%	77.00%	95.00%	88.70%	96.70%	54.30%	99.30%	78.80%	85.00%	76.20%	91.20%	53.80%	95.00%
MPP AOK	79.30%	98.70%	89.70%	99.30%	89.70%	100.00%	73.00%	94.00%	83.70%	96.30%	54.30%	99.70%	77.50%	87.50%	72.50%	92.50%	55.00%	96.20%
McNemar	×	×	×	×	✓	×	✓	×	✓	×	×	×	×	×	×	×	×	×

TABLE V

RESULTS FOR OUR MULTIPLE PARENTING APPROACH BY COMPARING FORESTS RECONSTRUCTED WITH E-AOB AGAINST GROUNDTRUTH FORESTS

		Metrics					
Set	**Algorithm**	**CR**	**CN**	**HP**	**HN**	**AP**	**AN**
DIP	**GT**	99.7%	99.7%	99.7%	99.7%	97.7%	100.0%
	E-AOB	81.0%	99.3%	91.0%	99.3%	86.0%	100.0%
McNemar		✓	×	✓	×	✓	×
PSB	**GT**	97.3%	97.3%	97.3%	97.3%	66.7%	100.0%
	E-AOB	77.0%	95.0%	88.7%	96.7%	54.3%	99.3%
McNemar		✓	✓	✓	×	✓	×
PRM	**GT**	96.2%	96.2%	90.0%	96.2%	60.0%	100.0%
	E-AOB	78.8%	85.0%	76.2%	91.2%	53.8%	95.0%
McNemar		✓	✓	✓	×	×	✓

TABLE VI

RESULTS FOR OUR MULTIPLE PARENTING APPROACH FOR THE PROFESSIONALLY SPLICING DATASET, DIVIDED IN THREE GROUPS ACCORDING TO HOW A HUMAN VIEWER PERCEIVES THE REALISM OF THE COMPOSED OBJECT

group	**case #**	**CR**	**CN**	**HP**	**HN**	**AP**	**AN**
all	**80**	**96.2%**	**96.2%**	**90.0%**	**96.2%**	**60.0%**	**100.0%**
easy	31	100.0%	100.0%	93.5%	100.0%	80.6%	100.0%
normal	27	100.0%	100.0%	88.9%	100.0%	59.3%	100.0%
hard	22	86.4%	86.4%	86.4%	86.4%	31.8%	100.0%

TABLE VII

J-LINKAGE'S AVERAGE NUMBER OF INPUT MATCHES AND COMPUTATION TIME (PRM TRAINING SET)

	Avg. Matches	**Avg. Time (in seconds)**
host	318.2	42.2
alien	138.6	18.8

errors due to inaccurate classification of trees and identification of parents. The most affected metric by using E-AOB to reconstruct the IPF is CR; the composition root quality. An incorrect CR can be due to two factors: the forest algorithm used found wrong roots for the forest (specifically the one of the composition tree) or the composition tree was incorrectly classified. From Table III, we see that the accuracy in finding the roots of the trees is of around 80%. In the E-AOB case, provided the proposed method was successful in classifying the trees, we should expect CR results around this percentage as well, albeit with a small variation, as the *roots* metric also takes into account other roots in the forest and the number of trees found. As seen by the obtained values of CR, in fact we have close to the expected 80%, showing that most of the impact in this metric is due to wrong roots found by the forest algorithms. A wrong composition root affects the identification of the alien and host parents, and thus the AP and HP metrics. Yet, we observe that the difference between AP and HP in GT and E-AOB at most of 12 percentage points, even though the differences in CR were much bigger, showing that a wrong composition root can still lead to correct classification of alien and host parents.

The superior results obtained in the scenario where the groundtruth IPF is available exhibit, for most metrics, statistical difference, emphasizing the importance of a correct forest reconstruction. Some metrics, such as AN or HN are harder to improve, as they already show very good results. The metrics related to identification of r_C, p_H and p_A, on the other hand, can still be improved exploring aspects of the method such as IPF reconstruction and detection of shared content among images.

Table VI shows the results for professionally-made compositions divided per difficult group and in a scenario of known IPFs (GT). Such forests were used to avoid effects from reconstruction algorithms in the results, as we intend to show only how properties of the composition, such as the size and blending of object, affect the multiple parenting results. The test cases division was based on their realism, as perceived by an human viewer, taking into account properties of the composition such as size and color of the composed object.

As expected, the accuracy, specially in alien parent identification decreases sharply in harder examples. Color, size, and blending properties, used to separate the set per difficult level affect the method in several ways. They decrease the accuracy of the forest reconstruction algorithms, by making it harder to separate the images belonging to the host and composition trees, as both share much content of the background. The classification is also affected, as detecting the shared content between alien and composition becomes significantly harder with smaller and more blended objects. Finally, the alien parent also becomes much harder to identify, due to the small number of pixels shared between it and the composition root. Still, the method shows some robustness in all but the AP metric, for all difficult groups. In this scenario, future work can focus on better alternatives for shared content detection and extraction, for improving the AP metric results.

In Table VII, we show J-Linkage's average computation time for comparisons between host and composition images and alien and composition images. The scenario studied was the identification of p_H and p_A, given the correct r_C and all candidate host and alien images, respectively. The computation times are divided by host and alien. Additionally, the average number of matches given as input to J-Linkage (after match selection) is also shown. The experiment was conducted on the training set of PRM.

The results give a glimpse of the dependence between J-Linkage's running time to the number of input matches. In the case of host images, the average number of matches is significantly greater, as host and composition images have much of their contents in common. Furthermore, it is important to highlight that in our preliminary work, the identification of the host parent was faster, as we used a pre-computed measure for this task. Using the Local Dissimilarity, which requires J-Linkage, for both alien and host parent identification, we sacrifice computational time to gain better results and a more general method to find p_H and p_A. Such advantages are specially meaningful if we want to apply our method to more complex scenarios. It is also important to highlight that every step involving the application of J-Linkage can take advantage of parallelization to mitigate the slower running time resulting from using the Local Dissimilarity. The current implementation runs on Matlab and does not have any optimization such as GPU paralellization and multiple threading.

C. Multiple Parenting Results With Missing Links and Unrelated Content

Real scenarios are likely to contain many images that are neither host, alien, or composition. The expanded test case set aims at simulating this, by adding, to each of the existing test cases in the PRM set, images that are unrelated to the multiple parenting generation process. Such additional images act as distractors, adding noise to several steps of the framework, such as the separation of groups and classification of trees. For example, if any of the unrelated images have similar contents to other images in the set, we might detect this as a shared content relationship, producing false positives in the tree classification process. Furthermore, real scenarios have another major complication: missing nodes. When dealing with NDIs obtained from the internet, it is practically impossible to acquire the complete phylogeny of an image. Thus, it is crucial that the proposed methods robustly deal with missing images.

We introduce three sets of test cases that take the aforementioned expanded test set, containing distractors, and further remove several randomly chosen nodes, plus one fixed choice node. This fixed choice can be any of the three nodes involved in the multiple parenting creation (Composition Root, Host Parent, or Alien Parent). The experiments in the additional sets seek to replicate, in a controlled setup, some of the main issues existent in uncontrolled scenarios, hence quantitatively measuring the robustness of our approach to them.

Table VIII shows the results for the expanded PRM set, using E-AOB and AOK for IPF reconstruction. The reduction of the performance when compared to the more controlled scenario of the PRM set (previous section), is caused by the noise introduced by the unrelated nodes in both the IPF reconstruction step and tree classification step. Adding unrelated nodes result in larger discrepancies in dissimilarity values between the images in the set, making it harder to separate hosts and compositions, which even though are not NDIs, have much of their content in common. Joining the host and composition IPTs has a direct effect into the classification of the composition and host IPTs. Additionally, unrelated nodes increase the chance of false positives in the shared content detection, also affecting IPT classification performance.

TABLE VIII

RESULTS FOR OUR MULTIPLE PARENTING APPROACH BY COMPARING THE TWO FOREST RECONSTRUCTION METHODS, AOK AND E-AOB, FOR THE EXPANDED PRM SET

	PRM-EXP					
Work	**CR**	**CN**	**HP**	**HN**	**AP**	**AN**
MPP E-AOB	53.80%	57.50%	43.80%	57.50%	30.00%	58.80%
MPP AOK	58.80%	66.20%	50.00%	65.00%	40.00%	67.50%
McNemar	×	×	×	×	×	×

In the expanded test cases, we can see a clear advantage when using AOK. In prior results, E-AOB has been shown to perform significantly better than AOK in cases with Semantically Similar Images [11], as well as in simpler multiple parenting scenarios (Table III), when taking into account classical Image Phylogeny evaluation metrics. However, as our results show, in less controlled scenarios with additional noise, the IPF reconstruction properties of AOK seem more suitable for multiple parenting. An advantage of AOK is the effect of the reconstruction parameter k (Table II). The IPF reconstructed by AOK is more stable to variations in this parameter, usually stabilizing for $k \leq 0$ and $k \geq 4$. E-AOB, however, requires a finer tuning of this parameter. Thus, it is easier to find an optimal value of k for AOK using a small training set than for E-AOB. In our experiments with the extended test cases, we used the same parameters trained in the PRM training set.

We now turn out attention to evaluating what happens when some nodes of the trees are missing. Table IX shows the results for this experiment. Missing nodes lead to larger transformation gaps between images from the same NDI group, which may result in the separation of images from such groups in two distinct IPTs when reconstructing the IPF. This separation usually impacts the composition tree recognition efficacy, as the analysis of shared content between trees will end up with many false positives among IPTs of the same NDI groups, but which ended up separated. As the shared content analysis is bounded by $O(t^2)$ (Section III-D), there is also a drop in efficiency.

In line with the results observed for the expanded PRM set, we notice that the AOK greedy approach is more suitable for less controlled scenarios than E-AOB. AOK is better at joining NDIs from the same group, even when their transformations are farther away (as it is common with missing nodes).

The metric's results maintain a certain consistency between the three sets. We observe that considering the two sets in which we forcefully remove the correct composition root (APin and HPin sets), the method still has solid substitute guesses for which node is the new composition root. When the composition root is missing, we evaluate as success if we point out as composition root any of the closest descendants of such node that were kept in the missing case scenario. Thus, it is still possible to have a good guess for the closest node to the original composition, even when this node is missing.

TABLE IX

RESULTS FOR OUR MULTIPLE PARENTING APPROACH BY COMPARING THE TWO FOREST RECONSTRUCTION METHODS, AOK AND E-AOB, FOR THE EXPANDED PRM SET WITH MISSING NODES

	CRin						HPin						APin					
Work	**CRS**	**CN**	**HPD**	**HN**	**APD**	**AN**	**CRS**	**CN**	**HPD**	**HN**	**APD**	**AN**	**CRS**	**CN**	**HPD**	**HN**	**APD**	**AN**
MPP E-AOB	43.80%	47.50%	1.8	51.20%	2.1	57.50%	42.50%	47.50%	2.2	55.00%	2.6	66.20%	40.00%	46.20%	2.6	50.00%	1.9	55.00%
MPP AOK	47.50%	58.80%	1.7	56.20%	2.0	71.20%	56.20%	61.30%	1.1	62.50%	3.0	72.50%	47.50%	61.30%	2.6	62.50%	2.0	68.80%
McNemar	×	✓	-	×	-	✓	✓	✓	-	×	-	×	×	✓	-	✓	-	✓

Fig. 9. Web scenario example images.

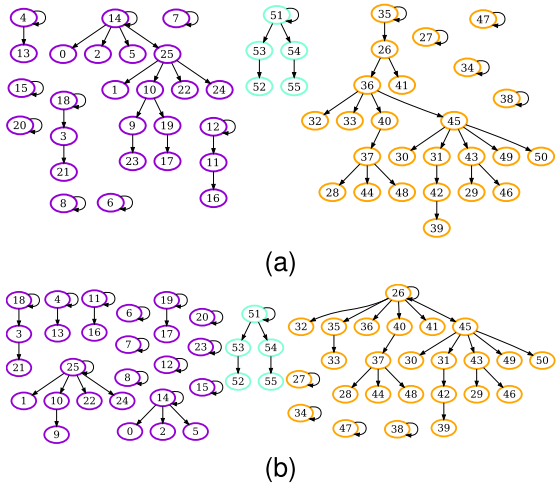

Fig. 10. IPF reconstruction of the web scenario using AOK (a) and E-AOB (b). Host nodes (purple) are numbered from 0 to 25, alien nodes (orange) are numbered from 26 to 50 and composition nodes (cyan) range from 51 to 55.

The distance metrics also show interesting results. Overall, they were consistent between all methods. In cases with the correct alien parent missing (CRin and HPin), for all methods, we are still able to guess a node that is close to the closest descendant of the eliminated alien parent, with distance of at most three edges. The resulting host parent is also close to the expected, with around two edges of distance.

D. Web Scenario Results

This section presents results for a test case containing host, alien and composition images obtained from the web (Section IV-B3). As we do not have a groundtruth for this case, our analysis is qualitative, and considers how reasonable the forest reconstruction, classification of trees and identification of multiple parents was. Fig. 9 depicts this scenario.

Fig. 10 shows the IPF reconstruction results for the web scenario, using AOK and E-AOB. The value of the parameter of k was chosen to ensure that we had at least one tree of each NDI groups (host, alien, and composition). For AOK, $k = 1.0$ and for E-AOB, $k = 0.8$. Both algorithms end up separating some of the NDIs from the same set into multiple different IPTs. This is more pronounced with E-AOB, in which the host nodes were separated in a total of 13 IPTs v. nine for AOK.

This separation is one of the biggest hurdles for the multiple parenting approach. It may affect the classification of the composition root, as the shared content relationships between IPTs from the same NDI group can result in ambiguities when classifying the composition root. It also becomes harder to classify the host and alien trees once we have the composition tree classified, as instead of only one candidate tree for each, we potentially have multiple trees that need to be classified into either host or alien.

Without loss of generality, for the remainder of this section, we used the tree reconstructed by AOK. In this test case, the identification of the composition root proceeded smoothly. For the majority of comparisons between nodes from the composition IPT, we observed some shared content relationships with the other trees, resulting in the correct classification of the tree rooted at node 51 as the composition IPT.

After classifying the composition IPT, the remaining trees must be classified into either host or alien. As some shared content relationships were observed between every other IPT and the composition IPT, all of them have to be classified. (If some of the trees did not have shared content, they could already be eliminated in this step). As we discussed in Section III-B, classification is done using the area of the shared region between the composition root and the root of a candidate tree. If this area is larger than a threshold $t\%$ of the composition root, we consider the candidate root to belong to a host IPT, otherwise, it belongs to an alien IPT. In the experiments, the threshold used was $t = 50\%$, although other values might be used at the risk of misclassifying alien IPTs as host, and vice-versa.

Using this procedure, only node 8 was misclassified, which is a very good result considering a real-world validation. The host image corresponding to node 8 is the image with the "LIE" overlay text, Fig. 9 depicts. This is perhaps one of the most transformed host images: it is grayscale and has severe noise, cropping, and a text overlay. The IPTs classified as host were the ones rooted at 4, 6, 7, 12, 14, 15, 18, 20, and the IPTs classified as alien were rooted at 8, 27, 34, 35, 38, 47.

Fig. 11. Multiple parents on web scenario. The alien parent had to be scaled down, as it is significantly bigger than the two other images.

Proceeding to parent identification, we consider the nodes of all the trees classified as host (and alien). Fig. 11 shows the identified parents. Both the host and alien parents are reasonable choices, as all their content are present in the composition, with no cropping. The chosen host parent also has a similar color in relation to the composition root, which is not true for all host images.

We believe the results for the web images is very promising, especially considering the complexity of the scenario. The method was able to find relationships between host, alien and composition images, even considering an imperfect separation of groups and an unknown number of missing nodes. Such results encourage us to wonder if it is possible to develop an image phylogeny method that also serves as a forgery detector and point out image relationships and hint at the composition process as a whole. However, there are still some challenges, such as the large number of trees from the same set of NDIs, and the possibility of having smaller, harder-to-detect objects sharing just a few pixels with the alien image.

We believe the results for the web images are promising, specially considering the complexity of the scenario. The method was able to find relationships between host, alien and composition images, even considering an imperfect separation of groups and an unknown number of missing nodes. Moreover, the experiments with unrelated images and missing links (Section VIII), which aimed at simulating some of the issues existent in uncontrolled scenarios, also showed some robustness of the proposed approach for the problem.

VI. Conclusion

The problem of Multiple Parenting Phylogeny arises when images are generated by combining the contents of two or more images. In this scenario, a relationship will exist between the source images, from where the content was extracted, and the new image created, which we refer to as a composition. Splicing compositions, created through pasting an object from a source images onto the background provided by another, were the used case study.

Here we employed a 3-step framework [12], which aims at reconstructing the phylogenies and multiple parenting relationships existent in a set of images containing NDIs of a composition and of its source images. To reconstruct the phylogenies, in addition to the heuristic Automatic Oriented Kruskal (AOK) algorithm used in our previous work, we applied an exact algorithm, the Extended Automatic Optimum Branching (E-AOB) [11]. We compared the two approaches, discussing their pros and cons. In general, we discovered that, although the literature shows that E-AOB outperforms AOK in general phylogeny problems, in the case of MPPs, there are some situations in which AOK is more appropriate. For instance, for real-world images collected from the internet whereby unrelated content may come along with near-duplicate images, AOK is more promising than E-AOB. The same happens when some nodes of the original trees are missing.

Moreover, we showed how using different keypoint detection parameters, allied with a new match selection policy, can improve the detection of shared content without significantly affecting efficiency. Finally, we also presented a new method of local comparison, combining our J-Linkage-based shared content extraction with the dissimilarity approach presented in [7] and [8] and its improvements [24]. Another novelty was the creation of a new dataset comprising professionally-made compositions, devised to appear more realistic to viewers, unlike the simpler compositions used in our previous work.

The multiple parenting results presented herein show promising results in identifying multiple parenting relationships in both image sets containing automatically created compositions and image sets containing professionally created compositions. For the alien parent identification, for instance, there is over 40% improvement for all composition types. When comparing with groundtruth forests, to simulate a scenario where we know the correct IPF, we discussed how much of the multiple parenting identification error arises from a bad forest reconstruction. In particular, there was a noticeable decrease in the identification of the original composition when using reconstructed forests. Yet, the other metrics remained reasonably stable, showing some robustness. We also showed a brief analysis of the results when applied to compositions with easier or harder to find objects. The used criteria was the size, color and blending of the object. As expected, there are drops in multiple parenting results for harder objects, although mostly on alien parent identification metrics.

Experiments with real-world images show the potential of the method for pointing out multiple parenting relations among images. With these results, we wonder if it is possible to develop an image phylogeny method that also serves as a forgery detector. By evolving the method to deal with large sets of uncontrolled images, which potentially contain many NDIs and some images with multiple parenting relationships, it might be possible to detect that an image is a forgery and also point out possible candidates that have been used to create it. Definitely, this is a research topic worth investigating.

The proposed method can also serve as a proxy for additional applications. Outside the realm of forensics, another potential field of application is studying how content changes and evolves within the internet. With image phylogeny, we have a glimpse about the individual evolution of images, and how they transform with time. However, there is no information about the creation of new content based on already existing content. For instance, an image that went viral is very likely to be used as a source to create new images, that can potentially go viral as well. The alien image in Figure 9, for example, was posted into a popular website, which prompted other users to copy it and create compositions such as the one depicted in Figure 9. Both aforementioned alien and composition images spread, generating NDI to be found in different websites. Multiple parenting phylogeny is

a potential tool to track and analyze the evolution of such images, by allowing us to identify the relationships between those different sets of images.

Looking ahead, there are still many challenges for applying multiple parenting to such scenarios and this paper serves as an invitation to the community to think about this problem more deeply. A correct separation of groups strongly affects the accuracy of the method, and is a very difficult task for similar images belonging to different groups, such as hosts and compositions. In our experiments, we found out that it can be very difficult to find a proper threshold that, at the same time, do not separate NDIs in different trees and also do not join similar but unrelated images (such as host and compositions, or semantically similar images) in the same tree. Additionally, compositions can be made in all forms and shapes, and sometimes an alien and composition image share only a few pixels worth of content. This turns detecting such objects, as well as separating host images from compositions, a daunting task. Besides those major challenges, generalization is of utmost importance for dealing with uncontrolled setups.

Generalization should be a major focus for future work. The improvements to the Local Dissimilarity measure helped steer our algorithm in this direction by unifying the measure used to identify host and alien parents. Notwithstanding, the proposed method is still dependent on the classification of the trees. By analyzing the relationships between pairs of trees, instead of trying to find an alien-composition-host classification, we can make the method more general and applicable to more complex scenarios. Such scenarios include cases with more than one composition, compositions created by combining more than two images, or test cases containing also completely unrelated images. Another branch of research could be focused on reducing the computation time when looking for shared content between images, specially when comparing images that share much of their contents. We can decrease such burden by optimizing the current steps of the framework, as well as exploring faster methods for image comparisons and parallelization of the framework.

REFERENCES

[1] H. Yang and J. Callan, "Near-duplicate detection by instance-level constrained clustering," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2006, pp. 421–428.

[2] M. Henzinger, "Finding near-duplicate Web pages: A large-scale evaluation of algorithms," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2006, pp. 284–291.

[3] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, "Near-duplicate keyframe identification with interest point matching and pattern learning," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 1037–1048, Aug. 2007.

[4] W.-L. Zhao and C.-W. Ngo, "Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 412–423, Feb. 2009.

[5] Y. Ke, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. ACM Multimedia*, 2004, pp. 869–876.

[6] L. Kennedy and S.-F. Chang, "Internet image archaeology: Automatically tracing the manipulation history of photographs on the Web," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 349–358.

[7] Z. Dias, A. Rocha, and S. Goldenstein, "First steps toward image phylogeny," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, Dec. 2010, pp. 1–6.

[8] Z. Dias, A. Rocha, and S. Goldenstein, "Image phylogeny by minimal spanning trees," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 774–788, Apr. 2012.

[9] A. De Rosa, F. Uccheddu, A. Costanzo, A. Piva, and M. Barni, "Exploring image dependencies: A new challenge in image forensics," *Proc. SPIE*, vol. 7541, p. 75410X, Jan. 2010.

[10] Z. Dias, S. Goldenstein, and A. Rocha, "Toward image phylogeny forests: Automatically recovering semantically similar image relationships," *Forensic Sci. Int.*, vol. 231, nos. 1–3, pp. 178–189, 2013.

[11] F. D. O. Costa, M. A. Oikawa, Z. Dias, S. Goldenstein, and A. R. de Rocha, "Image phylogeny forests reconstruction," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1533–1546, Oct. 2014.

[12] A. Oliveira *et al.*, "Multiple parenting identification in image phylogeny," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5347–5351.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[15] J. B. Kruskal, Jr., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, no. 1, pp. 48–50, 1956.

[16] Z. Dias, A. Rocha, and S. Goldenstein, "Exploring heuristic and optimum branching algorithms for image phylogeny," *J. Vis. Commun. Image Represent.*, vol. 24, no. 7, pp. 1124–1134, 2013.

[17] R. C. Prim, "Shortest connection networks and some generalizations," *Bell Syst. Tech. J.*, vol. 36, no. 6, pp. 1389–1401, 1957.

[18] Y. J. Chu and T. H. Liu, "On the shortest arborescence of a directed graph," *Sci. Sinica*, vol. 14, no. 10, pp. 1396–1400, 1965.

[19] J. Edmonds, "Optimum branchings," *J. Res. Nat. Bureau Standards*, vol. 71B, no. 4, pp. 233–240, 1967.

[20] F. Bock, "An algorithm to construct a minimum directed spanning tree in a directed network," *Develop. Oper. Res.*, vol. 1, no. 1, pp. 29–44, 1971.

[21] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, L. Del Tongo, and G. Serra, "Copy-move forgery detection and localization by means of robust clustering with J-Linkage," *Signal Process., Image Commun.*, vol. 28, no. 6, pp. 659–669, 2013.

[22] R. Toldo and A. Fusiello, "Robust multiple structures estimation with J-Linkage," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 537–547.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[24] F. de O. Costa, A. Oliveira, P. Ferrara, Z. Dias, S. Goldenstein, and A. Rocha, "New dissimilarity measures for image phylogeny reconstruction," Inst. Comput., Univ. Campinas, Campinas, Brazil, Tech. Rep. IC-15-07, Nov. 2015, pp. 1–20.

[25] (2015). *figshare*. [Online]. Available: http://http://figshare.com/

[26] (2015). *Github*. [Online]. Available: https://github.com/

[27] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM Special Interest Group Graph. Interact. Techn.*, 2003, pp. 313–318.

[28] (2015). *Imagemagick*. [Online]. Available: http://www.imagemagick.org/script/index.php

[29] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[30] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. Jul. 2001, pp. 416–423.

[31] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 416–431, Mar. 2006.

[32] K. McGuinness and N. E. O'Connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognit.*, vol. 43, no. 2, pp. 434–444, 2010.

[33] M. Marszatek and C. Schmid, "Accurate object localization with shape masks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[34] (2015). *Flickr*. [Online]. Available: https://www.flickr.com/

[35] (2015). *Wikimedia Commons*. [Online]. Available: http://commons.wikimedia.org/

[36] (2015). *Tineye*. [Online]. Available: http://www.imagemagick.org/script/index.php

[37] (2015). *Google Image Search*. [Online]. Available: https://www.google.com.br

Authors' photographs and biographies not available at the time of publication.