International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Near-Duplicate Image Analysis: Comprehensive Approaches to Image Phylogeny Tree Construction and Forensic Applications

Hemalata Mote[1*] and Sujata Kulkarni[2]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: The proliferation of digital technology enables unrestricted image creation and modification. These modified images often resurface on social media, creating a trail of near-duplicate images. This leads to challenges in tracking and verifying the origins and modifications of the image. In fields such as digital image forensics, news tracking services, and copyright enforcement, it is crucial to establish the connections between these modified images. An Image Phylogeny Tree (IPT) is created from a set of such altered images to map the sequence of changes at different levels. Image phylogeny organizes a series of logically similar images to highlight any modifications made to them. This paper investigates the construction of Image Phylogeny Trees (IPTs) to trace these relationships. It examines various algorithms—Oriented Kruskal (OK), Best Prim (BP), Optimum Branching (OB), Automatic Oriented Kruskal (AOK), and Automatic Optimum Branching (AOB)—for their efficiency in mapping the transformation sequence from an original image to its modified counterparts, providing a framework for systematic image alteration analysis.

Keywords: *Near-duplicates, Image Phylogeny Tree, Digital Image Forensics, Evolutionary Reconstruction, Algorithmic Evaluation, Dissimilarity calculation.*

1. Introduction

Examining and validating any available content is a difficult task because such content may have been changed, amended, or revised by various intermittent users. Because of the numerous changes that occur at various stages of content sharing or transfer, this scrutiny and validation become a major issue. When such content is presented as major legal evidence in litigation or investigations, this problem has serious consequences. Photographs and digital videos cannot be considered legal proof of evidence/occurrence due to the obvious ambiguity created around the origin and authentication of such contents. As a result, an established method of verification is required to protect the interests of the genuine copyright holder and legal owner of any such multimedia content creator [24]. Based on these considerations, significant research work has recently been dedicated to the forensic analysis of multimedia data. Because images are created by repeatedly applying photometric and geometric modifications, determining the sequence of modifications is a difficult problem [3]. To examine the history and modification process of digital images, the sequence of modifications is deemed as an Image Phylogeny Tree (IPT). An Image Phylogeny Tree that uses sequenced links to connect the root node (original image) to the child nodes (altered images).

IPT has the following applications:

Security: The document modification graph [10] can be used to obtain information about suspects' behavior and the history of modifications to any online document.

Forensics: If authentic documents are used for the forensic investigation rather than documents that are almost identical, better findings can be produced. Further, as they may be the original content creators, forensic investigators can concentrate on people connected to image recirculation around the root of the tree [11].

Copyright Imposition: Without the aid of active source control techniques (such as watermarking or fingerprinting), any infractions can be located [12].

News monitoring: Multimedia phylogeny is a useful technique for information collection. Close communication can shape perceptions over time and location, supplying important components of news tracing services [9] [15].

Image interconnection: IPT can be used to determine relationships between a complex set of photos metrically altered images. If all images come from a single tree or a different parent, IPT will be derived [1].

Detecting tampered images: Using variations between two images, IPT can deduce the trail of tampering done to the parent image [1].

Determining the parameters governing the transformations: The extent of transformation over two nodes of an edge can be analyzed using estimated parameters [1].

2. Discussions

Forgery detection, copyright infringement detection, and multimedia file matching all use the similarity function to detect and recognize near duplicates. In the multimedia

[1] *Sardar Patel Institute of Technology Mumbai, India.*
ORCID ID: 0000-0002-0714-3683
[12] *Sardar Patel Institute of Technology Mumbai, India.*
** Corresponding Author Email: hemalata.mote@spit.ac.in*

phylogeny idea, phylogeny trees are schematically structured via directed acyclic graphs (DAGs), and weights on edges explain the sequence of the alteration from parent to child. The dissimilarity function is used to generate edge weights that are less for identical objects and more for different objects. However, this dissimilarity in multimedia phylogeny may or may not constitute a distance in a metric space, based on the media being looked at. Depending on the media under consideration, this difference in multimedia phylogeny may or may not represent a distance in metric space [22]. A phylogeny forest is what happens when a number of trees, rather than just one, is used to depict the ancestral relationship of a group of nearly identical items. Semantically similar images are formed when a collection of near duplicates contains a number of distinct subsets. When the same location is used to take images or movies but different cameras are used, or when the same camera is used but different photo settings and positions are used [12] [8], these graphics are produced by combining many objects with semantically similar contents but coming from different sources. In the case of child pornography, it is crucial from a forensic perspective to identify all cameras and the holder in order for forensic experts to find the individual creators and original publishers of such content. If many cameras are used to capture the same scene. The multiple parenting phylogeny is an extension of the original work in image phylogeny forests; it focuses on establishing the correspondence between images that have nearly identical content (near-duplicates as well as those of apparently independent content) without any prior knowledge of how much content they share [7] [20].

Near-Duplicate Detection and Retrieval (NDDR) is a technique for locating fraud images [10] [11]. A suspect image may be a combination of several parts of images in some cases, i.e. parts of several images are combined to frame the resultant image [2]. The literature has looked into the hierarchical correspondence between a group of near-duplicate images. The results obtained have been mixed.

The donor photos in a composite can be located using provenance analysis methods [19]. Directionless phylogeny trees [19] are suited to these situations because figuring out the links is more significant than figuring out their direction.

The work of establishing ancestral relationships is carried out by IPTs (Image Phylogeny Trees), which also identify the root and derived nodes as well as the predecessors [11] [10]. It is assumed in IPT construction that all images are related in such a way that there is no independent node but only one root node [5] [6] [23]. Other works [10] [23] [21] create Image Phylogeny Forests by taking into account multiple root nodes.

The IPT design process is divided into two steps:

Step 1: Determine the degree of dissimilarity between two images using a pairwise asymmetric (dis)similarity measure. It also distinguishes between forward and reverses directional changes. It is possible to create a n x n matrix from a set of m near-duplicate images.

Step 2: A Tree Spanning Algorithm is used to derive the linkage between nodes and the hierarchical tree layout from the asymmetric matrix obtained in step 1.

Step 3: Final IPT is retrieved with local inheritance relationship correction.

The three steps of IPT design process are depicted in Fig. 1.

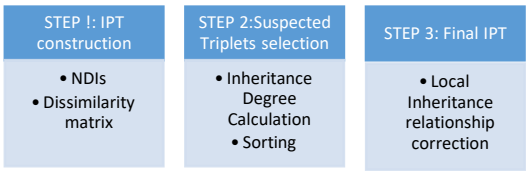

Fig 1. IPT Design Process

Dias et al., 2011 suggests that, the IPT is depicted as a minimal spanning tree (MST). In this case, the ancestral links are assumed to be correct edges. As a result, IPT is represented as a directed acyclic graph (DAG). A denoising auto encoder that replaces an approximation of the adjacency matrix corresponding to the underlying tree has recently been used to estimate an MST [20]. A greater number of studies focus on various geometric and pixel intensity-based transformations in images. Asymmetric measure calculation employs geometric registration, color channel normalization, and compression matching [13]. Melloni, A., et al. (2014) [17], employs the wavelet-based denoising technique, and [8], combines gradient estimation and mutual information techniques to obtain a better asymmetric measure. Melloni, A., et al. (2014) [17], is working to determine how many times an image has been edited. Thus, the age of an image is estimated by measuring the fitting of DCT coefficients and FDs statistics to a parametric model. Banerjee, S., and Ross, A. (2017) [1] recently performed photometric transformation modeling for iris images using simple linear and quadratic functions [23].

To estimate the transformation between a pair of images, Dias, Z., et al. (2010) [14] considered two different parametric models (a) Global Linear (GL) model and (b) Global Quadratic(GQ) model. The global models assume the application of the same set of transformation parameters to every pixel [23].

Global Linear (GL) Model:

The GL transformation model is denoted in equation (1), where a represents the multiplicative coefficient and b represents the bias or the offset term

$$T(\mathbf{I}|a, b) = a\mathbf{I} + b \quad (1)$$

Global Quadratic (GQ) Model:

This is considered to be a better model for the non-linearity's inherent to such transformations. It can be denoted as,

$$T(\mathbf{I}|a, b, c) = a\mathbf{I}^2 + b\mathbf{I} + c \quad (2)$$

Where, a, b and c represent the scalar coefficients of the transformation.

Banerjee, S., and Ross, A. (2019) [2] proposed photometric transformations for face images were modeled using Orthogonal polynomials and Gaussian Radial Basis Functions. To model the transformations and distinguish between forward and reverse transformations, a suitable set of basis functions must be chosen. There are several families of basis functions available, including the radial basis, polynomial basis, and wavelet basis. The associated weights, orthogonality values, and other characteristics of the basis functions within a family can vary Dissimilarity Calculation Techniques.

Compression matching: For image compression, JPEG compression parameters are used in this technique. This process produces artifacts on top of the target image.

Color matching: In this technique, by normalizing each channel by its corresponding channel's mean and standard deviation, the color of the source image is changed to match the color of the destination image [25].

Image registration, also known as geometric matching. There are various approaches discussed in the literature [26], and SURF (Speeded-Up Robust Features). Bay, H., et al. (2006) [4] computed key points in each pair of images to do image registration.

The dissimilarity calculation process [8] is shown below in Fig. 2 as follows:

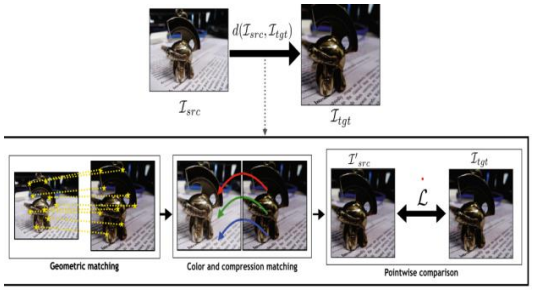

Fig 2. Histogram color matching

For a better color matching step, a histogram matching technique can be suggested [8]. This technique transforms the source image colors in such a way that their distribution acquires a form closer to the color distribution of the target image, by using the target image's color distribution information. Fig. 3 shows two examples of color matching algorithms.

To match the histograms of two images Ɪsrc and Ɪtgt, it is required to compute their histograms, Hsrc and Htgt and commute their Cumulative Distribution Function (CDF). For a grayscale image Ɪ, with L gray levels, the gray level I has the probability of

$$p\mathcal{I}(i) = ni/n,\ 0 \le I < L \quad (3)$$

Where n is the number of pixels in the image and ni is the number of pixels of gray value I in the histogram of image. The CDF of an image Ɪ is

$$CI(i) = \sum i\ k=0\ pI(k) \quad (4)$$

Once the mapping is found, each pixel with gray level. Each color channel of these images are treated independently, matching the histograms individually. Fig. 3 depicts the color distribution of the target image.

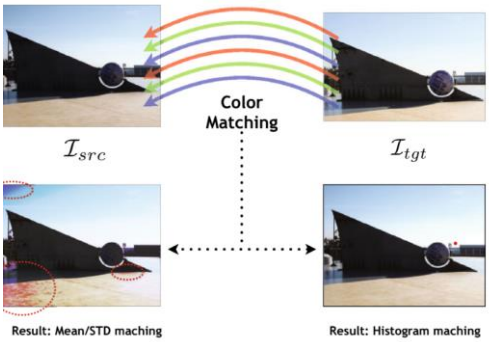

Fig 3. Colour distribution of the target image

Matching the colors of source image according to the colour distribution of the target image. Oliveira, A., et al. (2014) [23] shown the result of the color matching algorithm based on mean and standard deviation normalization presents undesirable artifacts that cannot be simply neglected, as can be noted in the marked regions of the picture. This problem is lessened when better color matching is performed through histogram analysis.

3. IPT Reconstruction Algorithms

The following reconstruction algorithms are used in [22] [14] [13] to reconstruct the Image phylogeny tree from near-duplicate media.

3.1. Types of Graphics

The dissimilarity matrix M was computed for each pair of images. It is a directed, full graph with weights along each edge. To reconstruct the phylogeny tree for dissimilarity matrix M, the authors in [14] [13] recommended using the classic Kruskal's minimum spanning tree algorithm to conform oriented graphs, which was dubbed Oriented Kruskal (OK). The assumption in this algorithm is that the root of the tree is not already known. The Algorithm aids in the discovery of the root and the construction of the oriented tree in only one execution. When there are n-1 edges for n near duplicate objects, the algorithm stops.

OK is a widely applied algorithm for finding a minimum spanning tree (MST) in a connected, weighted graph. In the

context of image phylogeny, it is employed to build a tree-like structure that corresponds to the relationships between near-duplicate images. The steps of the algorithm are explained below.

Step 1: Initialization.

The parent vector is initialized, which represents the tree structure. Every position Parenti identifies the parent of a node with ID i.

A total of n trees are initialized, each comprising a vertex representing an image.

Step 2: Iteration over dissimilarity matrix.

The algorithm is iterated over the dissimilarity matrix in order of dissimilarity, from minimum to maximum.

For every dissimilarity value at position (i, j) in the matrix, the following two tests are conducted:

Test 1 is to check if the endpoints i and j do not belong to the same tree. If they do not belong, then the next test is proceeded.

Test 2 is to check if j is the root of a tree. If it is, then the oriented edge (j→i) can be added to the forest.

If both the tests are passed, then the edge (j→i) is added to the forest, effectively linking the nodes i and j.

Step 3: Termination

All the nodes are connected by n-1 edges at the end of the algorithm, forming a tree that corresponds to the structure of changes of the original document on the basis of its near-duplications.

Thus, the resulting tree denotes the relationships between near-duplicate images, with the edges representing the dissimilarities or similarities between the images. The algorithm develops this tree by gradually linking the images based on their dissimilarity score, creating a hierarchical structure that captures the evolution or changes from the original image.

3.2. Best Prim (BP) Algorithm

Oriented prim is a new technique for reconstructing n different trees from the Dissimilarity Matrix M [10]. In this algorithm, each of the n nodes is only regarded as a root once. All possible roots will be assessed, and the image phylogenetic tree with the lowest reconstruction cost will be selected as the final one.

The BP algorithm is a variation of the Prim's algorithm, which is employed to identify a MST in a weighted, connected graph. In the context of image phylogeny, this algorithm is tailored to develop a phylogeny tree based on dissimilarity values between near-duplicate images.

The steps of the algorithm are explained below.

Step 1: Initialization.

A dissimilarity matrix M is received that captures the dissimilarity scores between each pair of n possible near-duplicate images.

For every possible root node r out of n, the following steps are performed:

Tr is initialized as an empty tree.

A priority queue Qr is initialized to keep track of the closest edge from the tree to the non-tree vertices.

Step 2: Running Oriented Prim (OP) algorithm as a sub-routine.

For every root node r, the OP algorithm is run as a subroutine:

The initial vertex r is chosen and is added to T_r.

The Q_r is updated with the dissimilarity values from r to all other vertices.

When T_r is seen not to cover all vertices, the following steps are performed:

The closest edge (u, v) is popped from Q_r and v is added to T_r.

The Q_r is updated by adding dissimilarity values from v to all vertices not in T_r.

Thus, the algorithm finally selects the tree with the lowest dissimilarity score as the image phylogeny tree, denoting the most likely evolutionary relationships among the near-duplicate images.

3.3. Optimum Branching (OB) Algorithm

To obtain the least branching, the Oriented Kruskal algorithm uses the smallest cost edge at each stage. In the Optimum Branching algorithm, the entire dissimilarity matrix infers the reconstruction of phylogeny [10]. One ideal global solution is used to improve the tree reconstruction evaluation matrix.

This algorithm aims to find an OB structure within a graph of image relationships. This structure denotes evolutionary or modification relationships between images with regards to image phylogeny. The steps of the algorithm are explained below.

Step 1: Assumption

A node from the graph is assumed as a potential root r.

Step 2: Finding minimum cost

For every node art from the assumed root r, the edge arriving at v is selected with the lowest cost. These selected edges denote the initial structure of branching.

Step 3: Circuit checking

The circuit is checked for closed loop edges known as

circuit. If no circuit is identified, then the next step is proceeded.

Or if a circuit is identified, the following steps are proceeded.

Step 4: Handling circuits

A dummy node vCi is created for representing the circuit.

The weights of edges connecting nodes outside the circuit (x) to nodes inside the circuit (y) are updated. This update takes into account the original edge weight w(x, y), adds the lowest weight of the edge within the circuit, and subtracts the edge weight of the edge arriving at y.

Step 5: Updating possible branches within the circuit

The weights of possible edges from nodes within the circuit to the outside are updated. These edges are part of the OB and must reflect the adjustments made because of the presence of the circuit.

The resultant structure represents an OB in the graph, considering the edges weights and handling any circuits if detected.

3.4. Automatic Oriental Kruskal (AOK)

AOK is an Oriented Kruskal algorithm extension that deals with images acquired automatically from various sources [10]. A Dissimilarity Matrix M is created in the algorithm for a set of semantically similar images. Each image is examined as a root of a tree. To connect trees, the algorithm considers a low-weight edge.

This algorithm is a special variant of Kruskal's algorithm for developing a forest of trees based on semantically similar images. The algorithm aims to control the inclusion of edges into the forest using a parameter AOK and statistical analysis of dissimilarity values. The steps of the algorithm are explained below.

Step 1: Calculation of input parameters. The algorithm takes the following 3 input parameters.

The number of semantically similar images n.

Dissimilarity matrix representing dissimilarities between images n×n.

A parameter calculated in advance, denoting to the number of standard deviations (SDs) used to limit the inclusion of edges in the forest γ_{AOK}.

Step 2: Variance-based Edge Inclusion.

The AOK algorithm monitors the variance of processed edges.

An edge is included to the forest only if its weight is lower than γ_{AOK} times the SD of the processed edges up to that point.

Step 3: Setting threshold for valid edges.

The parameter γ_{AOK} is related to a threshold point τ_{AOK} that chooses only the edges that belong to valid trees.

The threshold is defined as, $\tau_{AOK} = \mu_{AOK} + \gamma_{AOK} \times \sigma_{AOK}$

where μ_{AOK} is the average weight of edges and σ_{AOK} is the SD of edge weights already selected.

Step 4: Calculation of threshold.

γ_{AOK} is set to a predefined value on the basis of prior testing and analysis.

Step 5: Selection of edge.

The algorithm is iterated through the edges on the basis of dissimilarity scores in the M.

For each edge, the edge weight is compared with τ_{AOK}.

If the weight is less than τ_{AOK} , then the edge is added in the forest; otherwise, it is excluded.

Step 6: Forest construction.

This process is repeated until all valid edges have been processed and added in the forest.

The solution is a forest of trees representing semantically similar images, where edges are chosen based on their dissimilarity scores and τ_{AOK}.

Thus, the AOK algorithm adopts a statistical approach, using the Log-Normal distribution for guiding the inclusion of edges in the forest. By setting a threshold based on SDs and analyzing the dissimilarity scores, it ensures that only edges with dissimilarity scores within a certain statistical range are added.

3.5. Automatic Oriental Kruskal (AOK)

AOB considers all edges to achieve the least amount of branching. When some of the edges of a forest are removed, several independent partitions are formed. To improve the performance of AOB in [7], the author examined each partition separately. Edges that belong to current partitions are considered, so reconstruction results are improved with this re-execution where each tree is analyzed independently of the others. This results in E-AOB.

The AOB algorithm was developed to construct an OB forest based on semantically similar images. The steps of the AOB algorithm are explained below.

Step 1: Initialization.

The forest vector is initialized with n initial trees, each tree comprising a vertex that represents an image. The forest vector denotes the parent-child relationships in the trees.

Auxiliary parameters such as n edges (number of edges), x1 and x2 (variables for calculating the SD of accepted edges), and G (represents the graph constructed using M).

Step 2: Minimum Branching using OB algorithm.

Using the OB algorithm, the graph G is taken as input to calculate its minimum branching denoted by B.

Step 3: Edge sorting.

The edges of B are sorted based on non-decreasing order.

Step 4: Edge inclusion loop.

The following steps are iterated through the sorted edges of the B:

When the number of processed edges is greater than half of the edges of B, the SD is evaluated to limit edge inclusion based on AOK.

The number of accepted edges and their SDs are updated. The new edge in the forest vector is included, and the loop to process the next edge is repeated.

If the conditions are not satisfied, then the loop is exited.

The final result stored in the forest vector is returned as the solution. The goal is to intelligently choose edges for inclusion in the forest, aiming to achieve an OB while considering the SD to maintain stability and meaningful relationships.

E-AOB is extension of the Automatic Optimum Branching (AOB) algorithm. The steps of the E-AOB algorithm are explained below.

Step 1: Initialization.

The vector final Forest is initialized with n initial trees, each tree containing a vertex representing an image.

The variable G to represent the graph is constructed using the values from M.

Step 2: Construction of auxiliary graph for edge weights.

An auxiliary variable G' is constructed to store only the edges weights connected in the current forest.

The algorithm is iterated through the forest vector and the edge weights are extracted to represent subgraphs of the original graph G.

Step 3: Execution of AOB as sub-routine.

The AOB algorithm is executed using M and the edges represented by G'.

The OB is returned for each tree in the current forest.

Step 4: Updating forest with OB results.

The variable final Forest is updated with the new forest obtained from the AOB algorithm, representing the OB for each tree.

The updated resultant of final Forest is returned.

Step 5: Extended OB.

Based on the reconstructed forest, the E-AOB method is employed.

The M is updated to keep only the edges connected in the current forest.

The OB algorithm is executed again using the updated M.

The result is visually depicted, showcasing each tree with its corresponding color of sub-matrix, supporting in the interpretation, and understanding of the obtained OB structure for semantically similar images.

Table 1. shows a summary of recent state of art research papers, and Table 2. provides the comparative study on the different popular algorithms discussed in this section with respect to image phylogeny.

Table 1. Summary of recent state of art research papers

Reference	*Methodology*	*Dataset*	*Performance*
[3]	Parameter estimation using basis functions. To evaluate the reconstructed IPTs, use of Von Neumann graph entropy technique.	LFW CASIA iris V2 andV4 FVC 2000	Accuracy in Root identification ≈ 94% and Accuracy in IPT reconstruction ≈ 72%.
[6]	Estimation of Minimum Spanning Tree via denoising convolutional auto encoder.	UCID dataset	The percentage of correctly identified roots, edges, leaves, and ancestry links is used to evaluate a phylogenetic tree.
[19]	Provenance image filtering for provenance graph construction	NIST Professional Reddit	Contributed a novel clustering Algorithm. construction of own dataset from Reddit in-the-wild provenance cases for real-world provenance analysis.

[8]	Dissimilarity calculation using image gradients, MSE for standard pointwise comparison. Combining mutual information and gradient estimation techniques with a wavelet-based denoising technique	The Situation Room The Ellen DeGeneres selfie	A superior tree reconstruction after dissimilarity calculation pipeline due to color transformations.
[23]	By assessing the fitting of DCT coefficients and FDs statistics with respect to the parametric model, the age of the image is roughly determined.	UCID	Increase in IPT reconstruction accuracy and reduction in computational complexity.
[1]	For iris images, the Modelling of photometric transformations is done using local linear, global linear, and global quadratic functions.	CASIA iris V2 andV4	IPT-DAG reconstruction accuracy obtained with the Global Quadratic model is 71.30%
[2]	Modeling of photometric transformations for face images is done using polynomial and Radial Basis Functions.	LFW	Accuracy in IPT reconstruction is 67.53% and Accuracy in root identification is 87.88%.

Table 2. Comparative study on the popular image phylogeny algorithms

Algorithm	***Advantages***	***Disadvantages***	***Computational Complexity***
OK Algorithm	The computational complexity makes the algorithm suitable for processing larger number of images.	• The quality of the tree resulted by the algorithm is highly dependent on the accuracy and appropriateness of the dissimilarity metric used. • The algorithm treats edges as undirected edges, focusing essentially on capturing similarities and dissimilarities but not the directionality of the relationships.	$O(n^2 \log n)$
BP Algorithm	• The algorithm constructs an MST for every possible root and chooses the tree with the lowest dissimilarity sum. This ensures an optimal solution, capturing significant relationships while minimizing the overall dissimilarity score. • The algorithm is less likely to be influenced by the choice of root as it considers multiple roots to select the final optimal solution.	• The quality of the tree resulted by the algorithm is highly dependent on the accuracy and appropriateness of the dissimilarity metric used. • The computational complexity of the algorithm can still be a limitation for very large datasets.	$O(n^3)$
OB Algorithm	• The algorithm optimizes the branching structure within a graph by taking into account the edge costs, and handles circuit formations effectively. • The algorithm can process large graphs with multiple nodes and edges in a reasonable amount of time.	• The accuracy and informativeness of the resulting branching structure heavily relies on the edge costs or weights. • The computational complexity of the algorithm can still be a limitation for very large datasets.	$O(n^3)$
AOK Algorithm	• The algorithm uses Log-Normal distribution to guide the inclusion	• The accuracy and effectiveness of the algorithm heavily rely on the quality of the	$O(n^2 \log n)$

	of edges in the forest, thereby ensuring a statistically informed selection process, contributing to the accuracy and meaningfulness of the resulting forest. • The algorithm is specifically developed for semantically similar images, concentrating on capturing relationships based on dissimilarity metrics. • The computational complexity makes the algorithm suitable for processing larger number of images.	dissimilarity matrix M provided as input and the choice of parameter γAOK. • The computational complexity of the algorithm can still be a limitation for very large datasets.	
AOB Algorithm	• The algorithm considers statistical analysis to guide edge selection, ensuring a statistically informed inclusion procedure based on edge weights. • The algorithm develops an optimum branching forest on the basis of semantically similar images.	• The accuracy and effectiveness of the algorithm heavily rely on the quality of the dissimilarity matrix M provided as input and the choice of parameter γAOK. • The computational complexity of the algorithm can still be a limitation for very large datasets.	$O(n^3)$
E-AOB Algorithm	• The algorithm enhances the construction of the forest by using the information from an initial baseline solution provided by AOB algorithm. • This algorithm is adaptable to changes in the forest structure.	• The accuracy and efficiency of the algorithm is closely tied to the quality of the initial forest provided by the AOB algorithm. • The computational complexity of the algorithm can still be a limitation for very large datasets.	$O(n^3)$

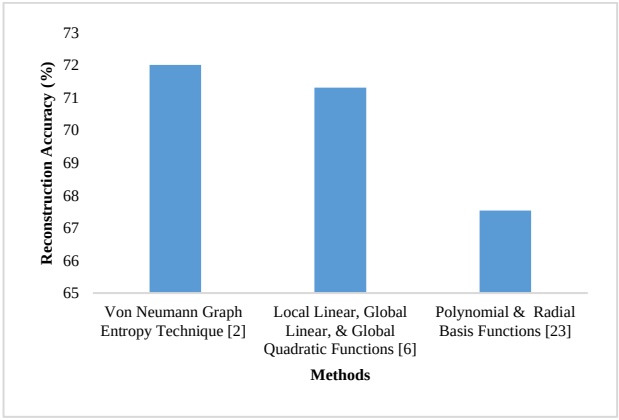

Fig 4. Comparison of reconstruction accuracy of different image phylogeny techniques

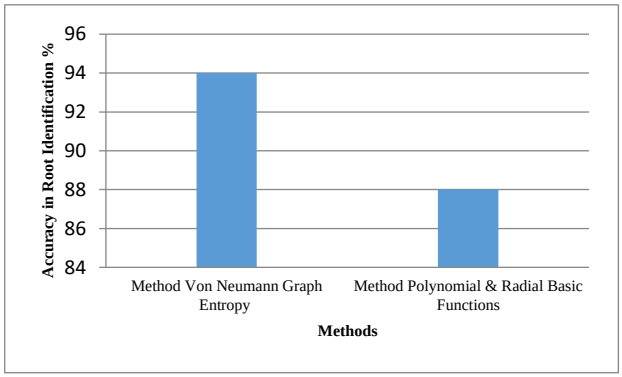

Fig 5. Comparison of accuracy in root identification of different image phylogeny techniques

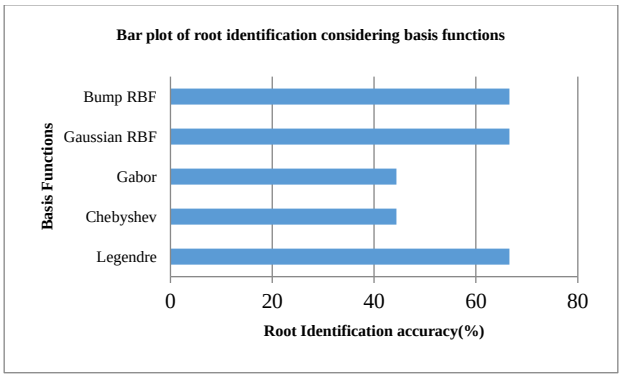

Fig 6. Comparison of root identification accuracy of different basis functions

Fig. 4 shows the comparison of the reconstruction accuracy obtained by different image phylogeny techniques. Von Neumann Graph Entropy technique shows higher reconstruction accuracy of 72% compared to local linear, global linear, global quadratic, polynomial and radial basis functions. Fig. 5 shows the comparison of root identification accuracy between Von Neumann Graph Entropy and polynomial and radial basis functions. It can be seen that the root identification accuracy of Von Neumann Graph Entropy is high up to 94%. Fig. 6 shows the comparison of root identification accuracies for different radial basis functions. The best performing radial basis function is found to be Bump, Gaussian, and

4. Challenges and Gaps

The survey of the literature identifies the following research gaps and challenges:

Algorithms used for IPT construction only perform local analysis.

Different transformations have an impact on the quality of reconstruction.

In the case of iris images, spurious edges are created, affecting reconstruction accuracy [1].

IPT reconstruction for a biometric trait, such as a fingerprint, is vulnerable to attack and has an impact on the acquisition rate.

Image acquisition methods are not specified.

Directed acyclic graphs could not be efficiently recovered.

The computational time for IPT is longer.

Photometric transformations are used to generate near duplicates. There are significant photometric transformations. Every transformation has a number of parameters. Again, each parameter has multiple value ranges. As a result, photometric transformation modeling becomes a difficult task. A multimodal biometric approach can be used for IPT reconstruction when only a few modalities are considered. It is difficult to obtain the biometric impression of the same person. Which will have an impact on performance parameters.

5. Proposed Methodology

From a near duplicate dataset, we can reconstruct IPT using two steps:

1) Creating an Asymmetric Dissimilarity matrix from a set of nearly identical images (i.e. Dissimilarity function)

2) Introducing a robust algorithm for constructing IPT from such a matrix.

We present a multimodal biometric authentication approach for reconstructing an Image Phylogeny Tree in the proposed work. We propose an optimal algorithm design that focuses on the use of biometric traits for Image Phylogeny Tree reconstruction. We want to use a set of basis functions to model any transformation. We also propose using image feature dimension reduction of the trained feature vector, which reduces memory size and thus computational time.

6. Conclusion

We examined various dissimilarity calculation methods as well as image phylogeny reconstruction algorithms. A group of photo shopped images is used to create an Image Phylogeny Tree (IPT), in order to map stepwise changes at different levels. Image phylogeny is a structure that traces the evolution of visually similar images over time displaying any type of image alteration. Pairwise analysis images were used in previous research on phylogeny trees to reconstruct IPT. By taking into account more potential basis functions for parameter modeling, the root node and IPT reconstruction accuracy can be increased.

Pairwise (dis)similarities are calculated to deal with multiple issues in computer vision, machine learning, and information retrieval to derive a degree of nearness among a set of objects. The asymmetry of this dissimilarity calculation can be used to determine parenthood relationships. The current approaches for video, audio, and text media types must be expanded. Furthermore, advancements in the calculation of dissimilarity between pairs of objects and in phylogeny reconstruction are required. Also, if IPT is generated for multimodal biometric images, how to perform dimension reduction and IPT computation time must be handled properly. Some of these issues must be addressed by research teams in the near future.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

Sujata Kulkarni: Conceptualization, Methodology Design, Writing-Reviewing and Editing

Hemalata Mote: Data curation, Writing-Original draft

preparation, Software, Validation., Field study

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Banerjee, S., and Ross, A. "Computing an image phylogeny tree from photo metrically modified iris images." In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 618-626, 2017, October. IEEE.

[2] Banerjee, S., and Ross, A. "Face phylogeny tree: Deducing relationships between near-duplicate face images using Legendre polynomials and radial basis functions." In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1-9, 2019, September. IEEE.

[3] Banerjee, S., and Ross, A. "Face phylogeny tree using basis functions." *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *2*(4), 310-325, 2020.

[4] Bay, H., Tuytelaars, T., and Gool, L. V. "Surf: Speeded up robust features." In *European conference on computer vision*, pp. 404-417, 2006, May. Springer, Berlin, Heidelberg.

[5] Bestagini, P., Tagliasacchi, M., and Tubaro, S. "Image phylogeny tree reconstruction based on region selection." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2059-2063, 2016, March. IEEE.

[6] Castelletto, R., Milani, S., and Bestagini, P. "Phylogenetic minimum spanning tree reconstruction using auto encoders." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2817-2821, 2020, May. IEEE.

[7] Costa, F. O., Oikawa, M., Dias, Z., Goldenstein, S., and Rocha, A. "Image phylogeny forests reconstruction." IEEE Transactions on Information Forensics and Security, 9(10):1533-1546, 2014.

[8] Costa, F., Oliveira, A., Ferrara, P., Dias, Z., Goldenstein, S., and Rocha, A. "New dissimilarity measures for image phylogeny reconstruction." *Pattern Analysis and Applications*, *20*(4), 1289-1305, 2017.

[9] De Rosa, A., Uccheddu, F., Costanzo, A., Piva, A., and Barni, M. "Exploring image dependencies: a new challenge in image forensics." In *Media forensics and security II*, Vol. 7541, pp. 337-348, 2010, January. SPIE.

[10] Dias, Z., Goldenstein, S., and Rocha, A. "Exploring heuristic and optimum branching algorithms for image phylogeny." *Journal of Visual Communication and Image Representation*, *24*(7), 1124-1134, 2013.

[11] Dias, Z., Goldenstein, S., and Rocha, A. "Large-scale image phylogeny: Tracing image ancestral relationships." *IEEE Multimedia*, *20*(3), 58-70, 2013.

[12] Dias, Z., Goldenstein, S., and Rocha, A. "Toward image phylogeny forests: Automatically recovering semantically similar image relationships." *Forensic science international*, *231*(1-3), 178-189, 2013.

[13] Dias, Z., Rocha, A., and Goldenstein, S. "Image phylogeny by minimal spanning trees." *IEEE Transactions on Information Forensics and Security*, *7*(2), 774-788, 2011.

[14] Dias, Z., Rocha, A., and Goldenstein, S. "First steps towards image phylogeny." *In IEEE International Workshop on Information Forensics Security*, pp. 1-6, 2010.

[15] Kennedy, L., and Chang, S. F. "Internet image archaeology: Automatically tracing the manipulation history of photographs on the web." In *Proceedings of the 16th ACM international conference on Multimedia*, pp. 349-358, 2008, October.

[16] Le Philippe, N., Puech, W., and Fiorio, C. "Phylogeny of JPEG images by ancestor estimation using missing markers on image pairs." In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1-6, 2016, December. IEEE.

[17] Melloni, A., Bestagini, P., Milani, S., Tagliasacchi, M., Rocha, A., and Tubaro, S. "Image phylogeny through dissimilarity metrics fusion." In *2014 5th European Workshop on Visual Information Processing (EUVIP)*, pp. 1-6, 2014, December. IEEE.

[18] Milani, S., Fontana, M., Bestagini, P., and Tubaro, S. "Phylogenetic analysis of near-duplicate images using processing age metrics." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2054-2058, 2016, March. IEEE.

[19] Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K. W., and Scheirer, W. J. "Image provenance analysis at scale." *IEEE Transactions on Image Processing*, *27*(12), 6109-6123, 2018.

[20] Oikawa, M. A., Dias, Z., de Rezende Rocha, A., and Goldenstein, S. "Manifold learning and spectral clustering for image phylogeny forests." *IEEE Transactions on Information Forensics and Security*, *11*(1), 5-18, 2015.

[21] Oikawa, M. A., Dias, Z., de Rezende Rocha, A., and Goldenstein, S. "Manifold learning and spectral

clustering for image phylogeny forests." *IEEE Transactions on Information Forensics and Security*, *11*(1), 5-18, 2015.

[22] Oikawa, Marina A., et al. "Distances in multimedia phylogeny." *International Transactions in Operational Research* 23.5: pp. 921-946, 2016.

[23] Oliveira, A., Ferrara, P., De Rosa, A., Piva, A., Barni, M., Goldenstein, S., ... and Rocha, A. "Multiple parenting identification in image phylogeny." In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5347-5351, 2014, October. IEEE.

[24] P. Bestagini et al., "An overview on video forensics," 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 1229-1233, 2012.

[25] Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. "Color transfer between images." *IEEE Computer graphics and applications*, *21*(5), 34-41, 2001.

[26] Zitova, B., and Flusser, J. "Image registration methods: a survey." *Image and vision computing*, *21*(11), 977-1000, 2003.