Phylogeny of JPEG images by ancestor estimation using missing markers on image pairs

Noé Le Philippe, William Puech and Christophie Fiorio

LIRMM Laboratory, CNRS, University of Montpellier, France

December 8, 2016

What is image phylogeny

Definition

"In biology, phylogenetics is the study of evolutionary history and relationships among individuals or group of organisms" — Wikipedia

What is image phylogeny

Definition

"In biology, phylogenetics is the study of evolutionary history and relationships among individuals or group of organisms" — Wikipedia

For images

The study of evolutionary history and relationships among images

What is image evolution

Near duplicates[1]

$$I_{n+1} = T(I_n),\ T \in \mathcal{T}$$

[1] Alexis Joly, Olivier Buisson, and Carl Frélicot. "Content-based copy retrieval using distortion-based probabilistic similarity search". In: *Multimedia, IEEE Transactions on* 9.2 (2007), pp. 293–306.

What is image evolution

Near duplicates[1]

$I_{n+1} = T(I_n),\ T \in \mathcal{T}$

Tolerated transformations

- transformations in $\mathcal{T}$
- magnitude of the transformation

[1] Alexis Joly, Olivier Buisson, and Carl Frélicot. "Content-based copy retrieval using distortion-based probabilistic similarity search". In: *Multimedia, IEEE Transactions on* 9.2 (2007), pp. 293–306.

What is image evolution

Near duplicates[1]

$I_{n+1} = T(I_n),\ T \in \mathcal{T}$

Tolerated transformations

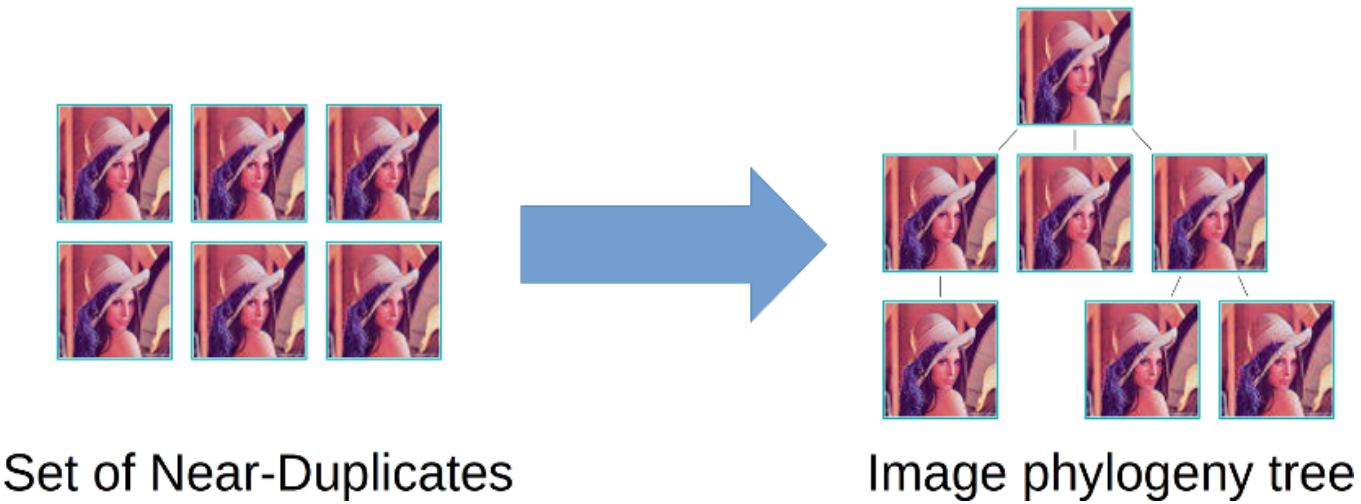
- transformations in $\mathcal{T}$
- magnitude of the transformation

Our goal

Compute an image phylogeny tree of near duplicates where $\mathcal{T} = \{lossy\ compression\}$

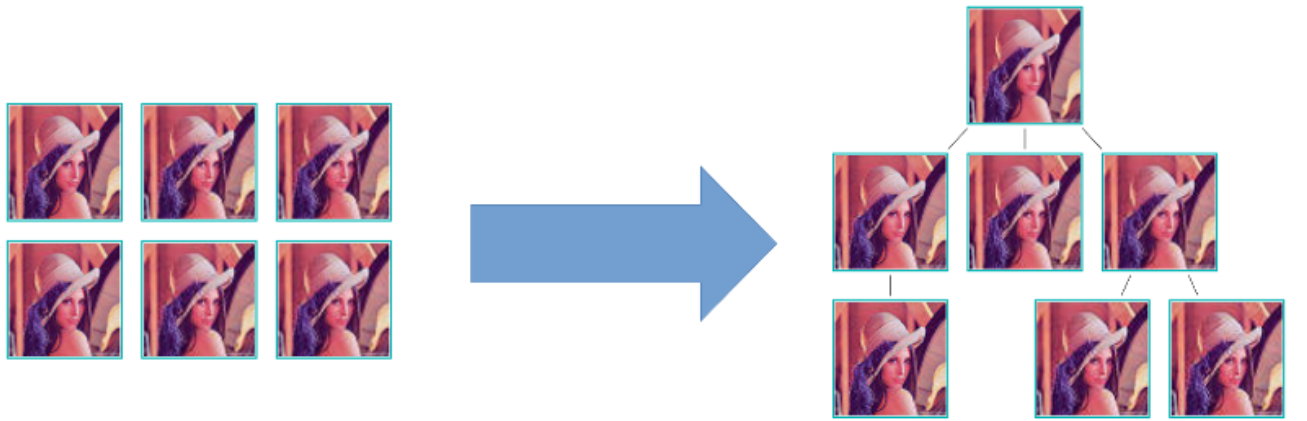
[1] Alexis Joly, Olivier Buisson, and Carl Frélicot. "Content-based copy retrieval using distortion-based probabilistic similarity search". In: *Multimedia, IEEE Transactions on* 9.2 (2007), pp. 293–306.

Image phylogeny tree



Two parts during phylogeny tree reconstruction :

Image phylogeny tree



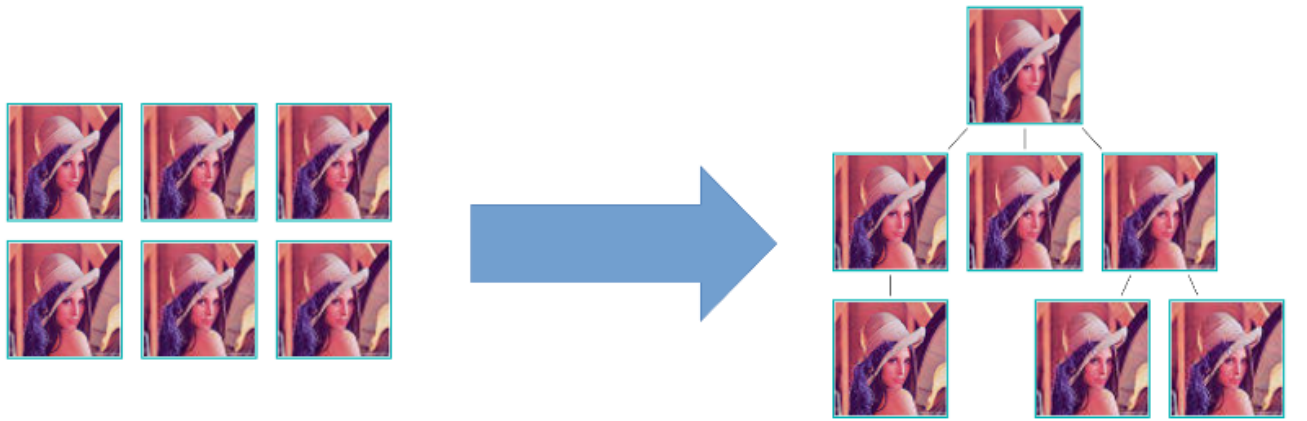
Set of Near-Duplicates

Image phylogeny tree

Two parts during phylogeny tree reconstruction :

- Identify the root

Image phylogeny tree



Set of Near-Duplicates

Image phylogeny tree

Two parts during phylogeny tree reconstruction :

- Identify the root
- Estimate the rest of the tree

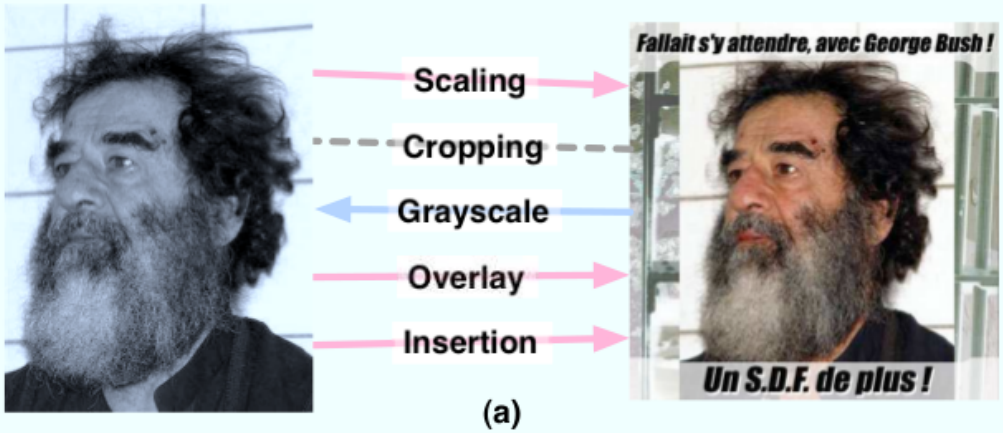
Estimating the image phylogeny tree

Visual Migration Map[2]

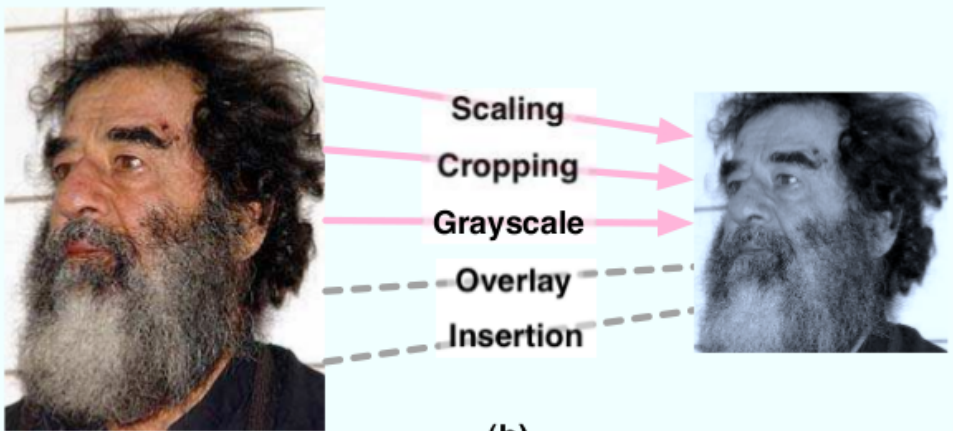
- Image manipulations are directional
- Parent-child relationship if manipulation directions are consistent
- Tree reconstruction using the longest path

[2] Lyndon Kennedy and Shih-Fu Chang. "Internet image archaeology: automatically tracing the manipulation history of photographs on the web". In: *Proceedings of the 16th ACM international conference on Multimedia*. ACM. 2008, pp. 349–358.

Estimating the image phylogeny tree



(a)
Inconsistent directions from individual manipulations.
(Neither image is parent)



(b)
Consistent directions from individual manipulations.
(Left image is parent of right)

Estimating the image phylogeny tree

Image phylogeny tree [3] [4]

- Extraction of a *dissimilarity matrix*
- Tree reconstruction using a minimum spanning tree algorithm (e.g. Kruskal)

---

[2] Zanoni Dias, Anderson Rocha, and Siome Goldenstein. "First steps toward image phylogeny". In: *Information Forensics and Security (WIFS), 2010 IEEE International Workshop on*. IEEE. 2010, pp. 1–6.

[3] Zanoni Dias, Anderson Rocha, and Siome Goldenstein. "Image phylogeny by minimal spanning trees". In: *Information Forensics and Security, IEEE Transactions on* 7.2 (2012), pp. 774–788.

Table of content

1 Introduction

2 Our method

3 Results

4 Conclusion

Goal

Reduce a complex image phylogeny tree reconstruction to a simple **parent-child relationship negation**

Solution

Binary decision between two images : *"Can this image be an ancestor of the other one?"*

Our method

Marker

A marker is a local or global feature extracted from the image. This marker shows that a particular transformation was applied to the image. This marker is passed on to the image child

Our method

Marker

A marker is a local or global feature extracted from the image. This marker shows that a particular transformation was applied to the image. This marker is passed on to the image child

Negation function

Let $f(I_m, I_n)$ be a function that for every image pair (I_m, I_n) detects every time there is one a marker visible in the image I_m and not in its potential child I_n, thus proving that I_m is not an ancestor of I_n

Our method

Marker

A marker is a local or global feature extracted from the image. This marker shows that a particular transformation was applied to the image. This marker is passed on to the image child

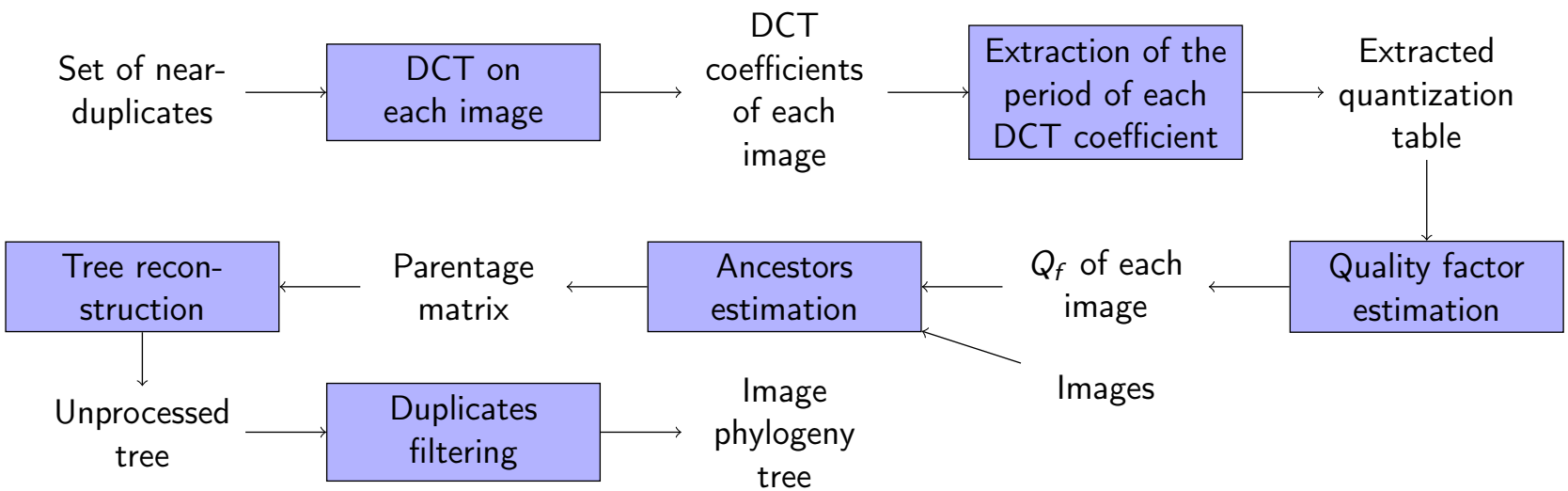
Negation function

Let $f(I_m, I_n)$ be a function that for every image pair (I_m, I_n) detects every time there is one a marker visible in the image I_m and not in its potential child I_n, thus proving that I_m is not an ancestor of I_n

Theorem

For all image pairs (I_m, I_n) in a set of near-duplicate images, if there is no marker proving that I_m is not an ancestor of I_n then there is a parent-child relationship between I_m and I_n, $I_m \rightarrow I_n$, with $m < n$

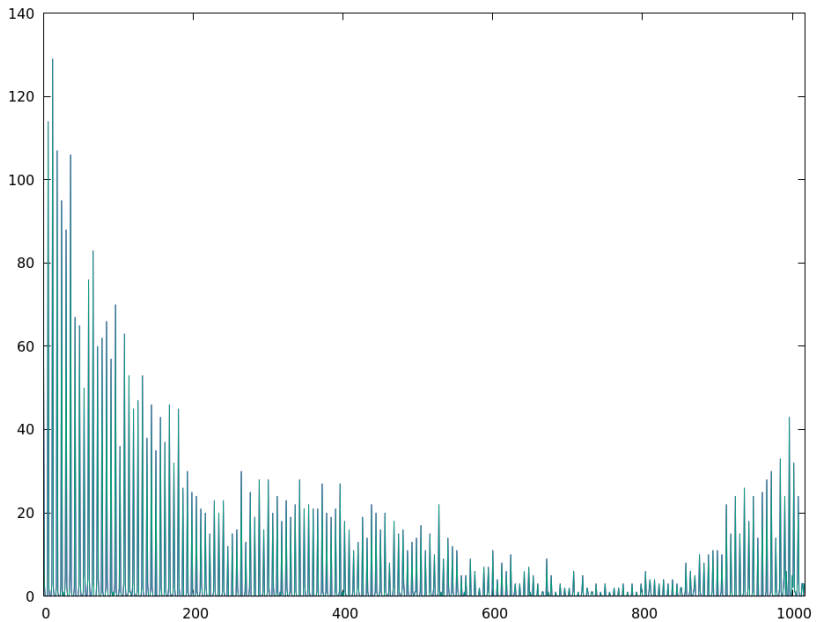
Diagram of our method

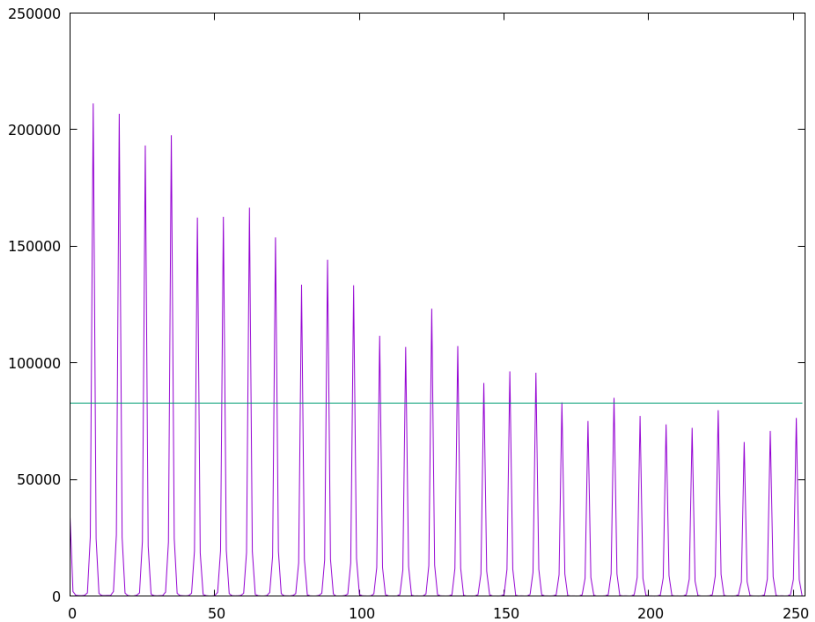


$\widehat{q}(u, v)$ estimation

How to get $\widehat{q}(u, v)$

- Values of a quantized signal gather around the quantization step
- Delta between each peak of the autocorrelation of every DCT coefficient





$\widehat{q}(u, v)$ estimation

9	6	5	9	13	22	22	30
6	6	8	10	14	16	30	
8	7	9	13	20	31		
8	9	12	19	28			
10	12	-1	-1				
12	13	30					
28	31						

Figure: Example of a quantization table $\widehat{q}(u, v)$ returned by the period estimation

Limited to the 35 first coefficients

Quality factor estimation

Primary estimation

$$Q_f = argmin \sqrt{\sum_{i=0}^{7} \sum_{j=0}^{7} |\widehat{q}(i,j) - q_n(i,j)|^2}$$

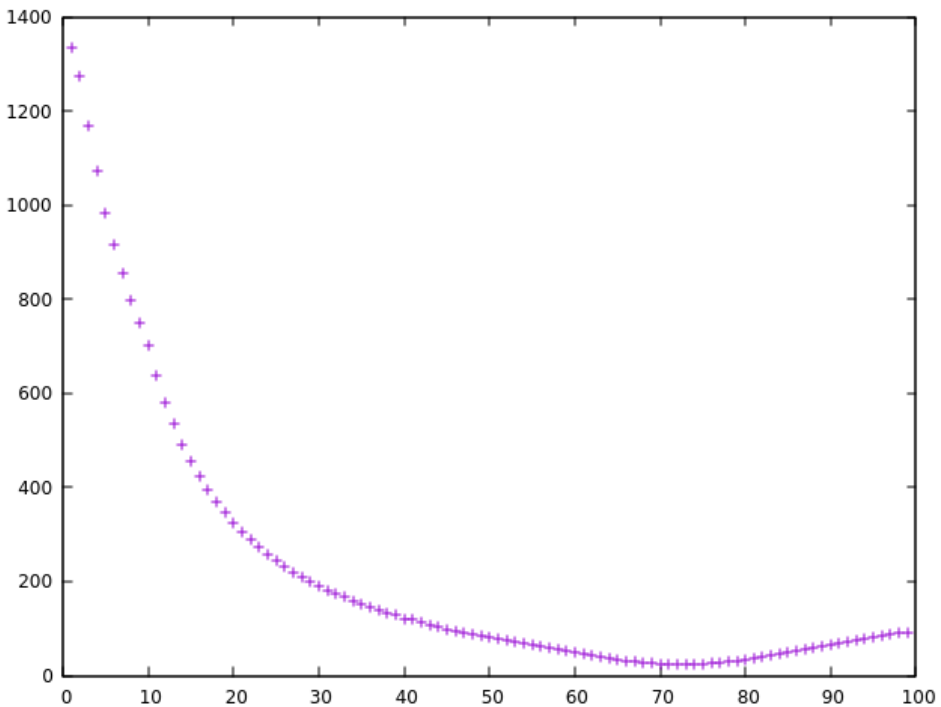
where q_n is a JPEG quantization table and $1 \leq n \leq 100$

Secondary estimation

- ***if*** $Q_f < 50$ $Q_s = 5000/Q_f$ ***else*** $Q_s = 200 - (Q_f \times 2)$ (1)
- $q(u,v) = \frac{(q_{50}(u,v) \times Q_s) + 50}{100}$ *with* $1 \leq q(u,v) \leq 255$ (2)

where q_{50} is the reference JPEG quantization table for $Q_f = 50$.

Quality factor estimation - example



(a) Plot of the distances between $\widehat{q}(u, v)$ and $q_n(u, v)$

16	12	14	14	18	24	49	72
11	12	13	17	22	35	64	92
10	14	16	22	37	55	78	95
16	19	24	29	56	64	87	98
24	26	40	51	68	81	103	112
40	58	57	87	109	104	121	100
51	60	69	80	103	113	120	103
61	55	56	62	77	92	101	99

8	6	7	7	9	12	25	36
6	6	7	9	11	18	32	46
5	7	8	11	19	28	39	48
8	10	12	15	28	32	44	49
12	13	20	26	34	41	52	56
20	29	29	44	55	52	61	50
26	30	35	40	52	57	60	52
31	28	28	31	39	46	51	50

Quality factor estimation - example

Example for $u = 0$ and $v = 0$

Data

$\widehat{q}(0,0) = 9, \;\; q_{50}(0,0) = 16$

Using equation (2)

$9 = \frac{16 \times Q_s + 50}{100}, \quad Q_S = \frac{9 \times 100 - 50}{16} = 53.125$

Using equation (1)

$53.125 = 200 - (Q_f \times 2), \quad Q_f = \frac{200 - 53.125}{2} = 73.4375$

This process is repeated for every element in $\widehat{q}(u, v)$ and averaged to obtain an accurate estimation of Q_f

Ancestors estimation

JPEG compression is deterministic

JPEG compression is not transitive

The parent is one compression away from its children

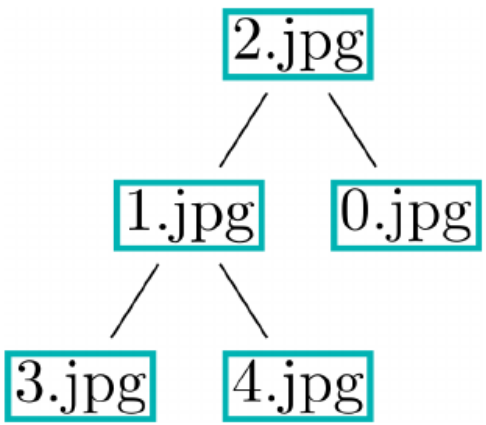
Filtering images that can't be an ancestory

Binary decision

Tree reconstruction

Binary matrix of size $n \times n$

Tree construction from the matrix



(a) Phylogeny tree

-	I_0	I_1	I_2	I_3	I_4
I_0	-	0	0	0	0
I_1	0	-	0	1	1
I_2	1	1	-	1	1
I_3	0	0	0	-	0
I_4	0	0	0	0	-

(b) Parentage matrix

Table of content

1 Introduction

2 Our method

3 Results

4 Conclusion

On full trees

A very good estimation of Q_f

Metrics close to 100%

A decreased accuracy with bigger trees

On full trees

Metric \ Dataset	**15 images**	**25 images**	**50 images**
Average Q_f estimation error	0.42	0.64	0.83
roots	95.83	88.88	84.72
edges	99.70	99.24	98.97
leaves	99.59	99.15	98.74
ancestry	99.44	96.88	96.96

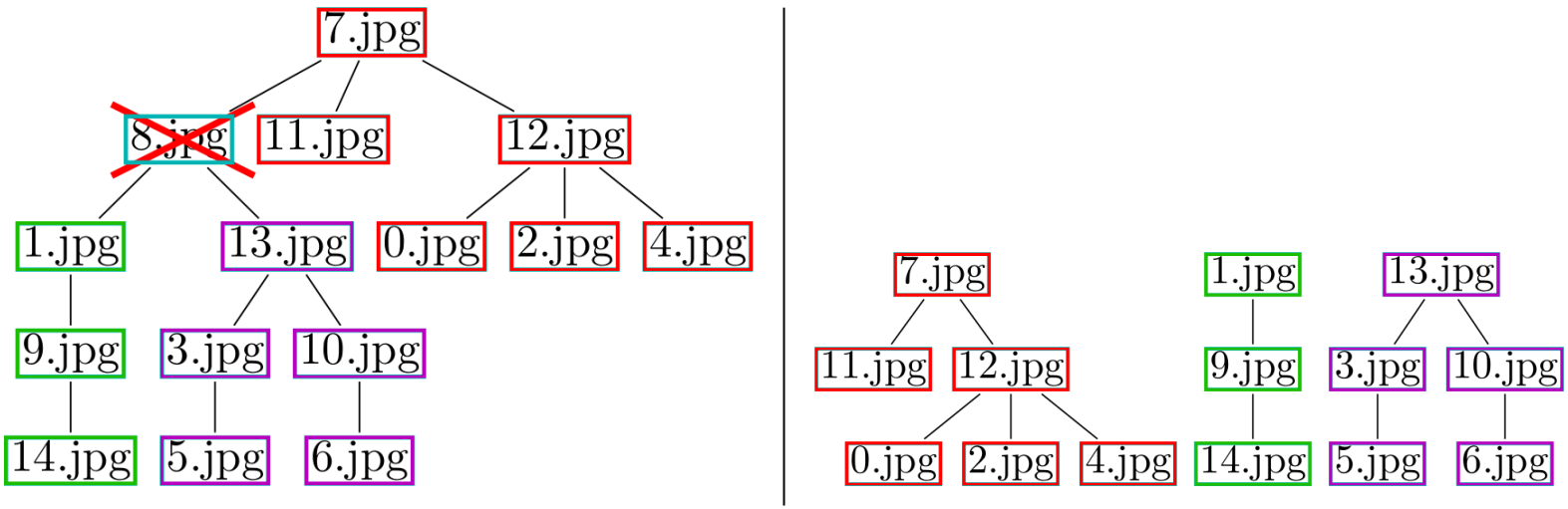
Trees with a missing image

Bad root identification

Good estimation of the rest of the tree

Our method only detects parents

Trees with a missing image



Trees with color images

No change in our implementation

We used the luma component

Good results

Trees with color images

Metric \ Dataset	15 images	25 images	50 images
Average Q_f estimation error	1.15	1.29	1.42
roots	93.94	81.82	87.88
edges	99.35	98.61	99.38
leaves	99.62	98.66	99.78
ancestry	98.79	94.13	98.82

Table of content

1 Introduction

2 Our method

3 Results

4 Conclusion

Conclusion - Perspectives

Conclusion

- A promising method with good results
- Limited to parents

Perspectives

- Explore new distortions
- Discover new markers

Questions ?