La phylogénie des images dans les réseaux sociaux

Noé LE PHILIPPE

Équipe ICAR - William Puech

16 juin 2016

Sommaire

Le sujet de stage

Le sujet

La phylogénie des images dans les réseaux sociaux

Le sujet de stage

Le sujet

La phylogénie des images dans les réseaux sociaux

Définition

"La phylogenèse ou phylogénie est l'étude des relations de parenté entre êtres vivants." — Wikipedia

Les applications

Réduire le nombre de versions de la même image pour optimiser l'espace de stockage

Les applications

Réduire le nombre de versions de la même image pour optimiser l'espace de stockage

Suivre l'évolution et la diffusion d'images sur les réseaux sociaux

Les applications

Réduire le nombre de versions de la même image pour optimiser l'espace de stockage

Suivre l'évolution et la diffusion d'images sur les réseaux sociaux

Détecter l'altération d'images

Définitions

Near-Duplicate Image (NDI) [1]

Une image I_n est le near-duplicate d'une image I_m si :

$$I_n = T(I_m), T \in \mathcal{T}$$

où $\mathcal{T}$ est un ensemble de transformations autorisées

Dans le cas général,

$$\mathcal{T} = \{resampling, cropping, affine\ warping, color\ changing, lossy\ compression\}$$

mais dans le cadre du stage, $\mathcal{T} = \{lossy\ compression\}$

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

Définitions

Image Phylogeny Tree (IPT)

C'est l'arbre retraçant la parenté des images

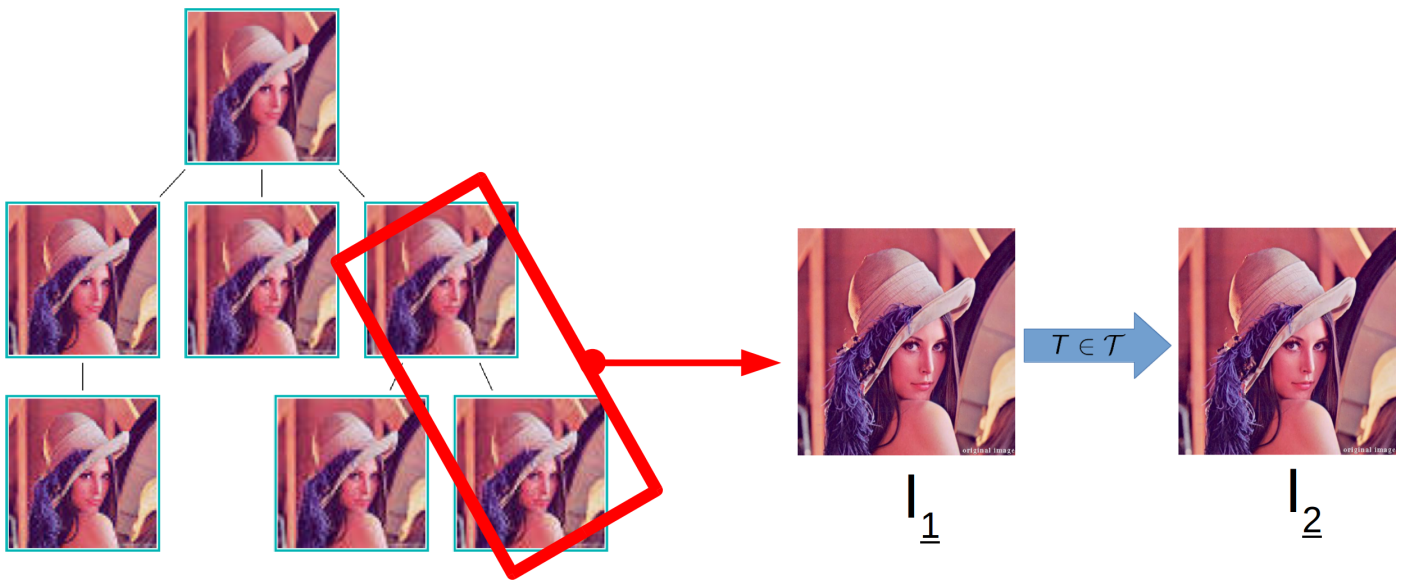
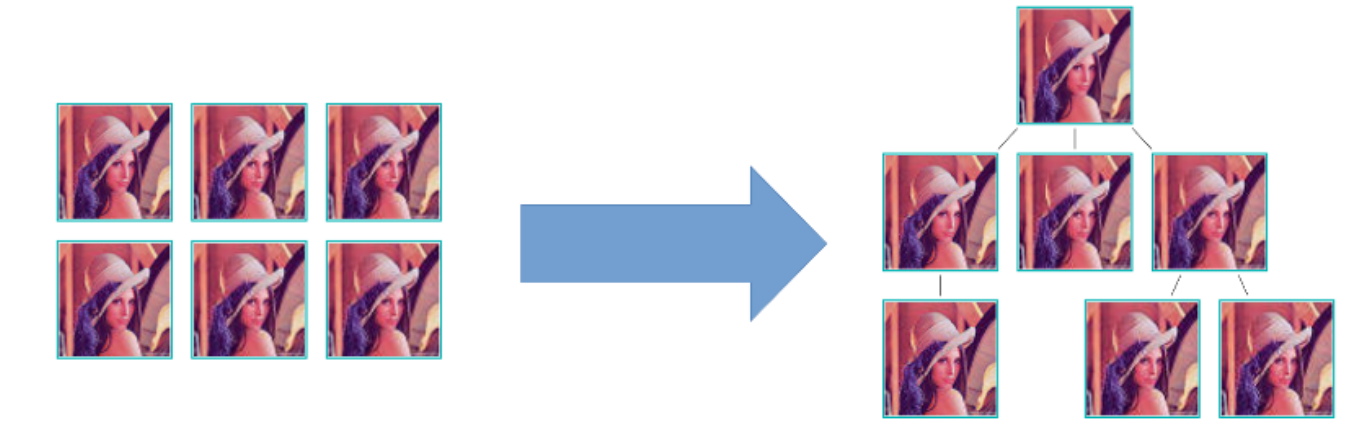
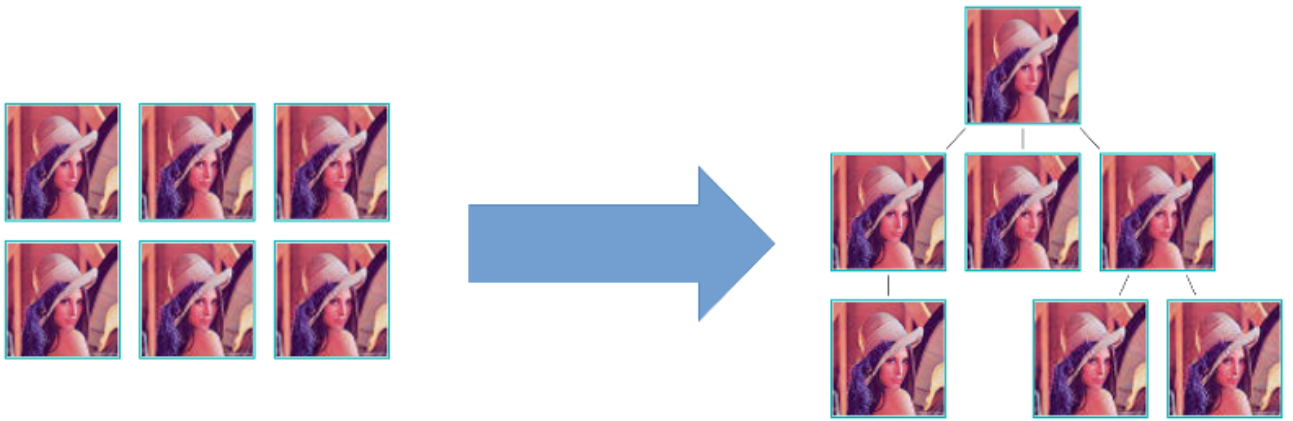


Image phylogeny tree



Deux parties importantes lors de la reconstruction de l'arbre phylogénétique :

Image phylogeny tree



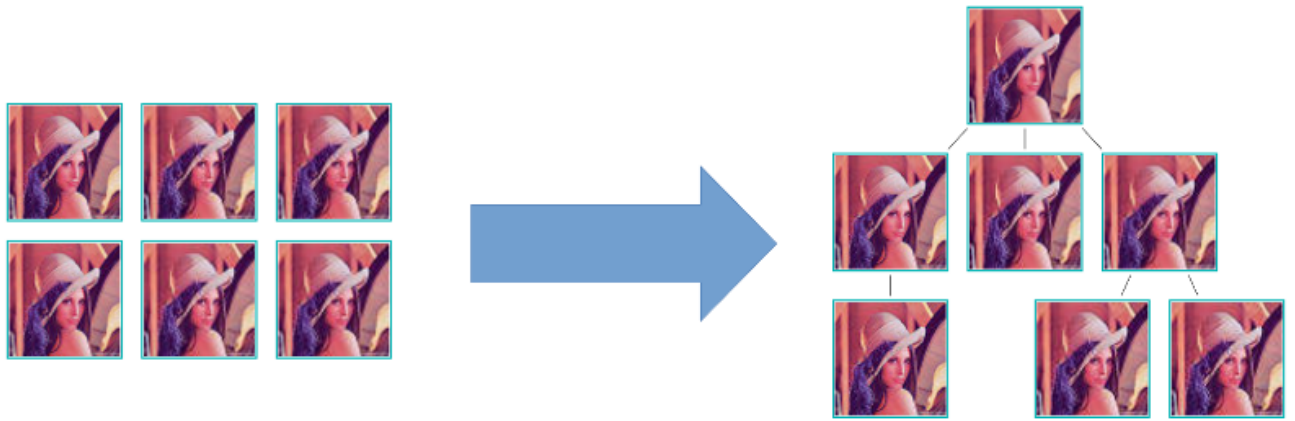
Set of Near-Duplicates

Image phylogeny tree

Deux parties importantes lors de la reconstruction de l'arbre phylogénétique :

- Correctement identifier la racine

Image phylogeny tree



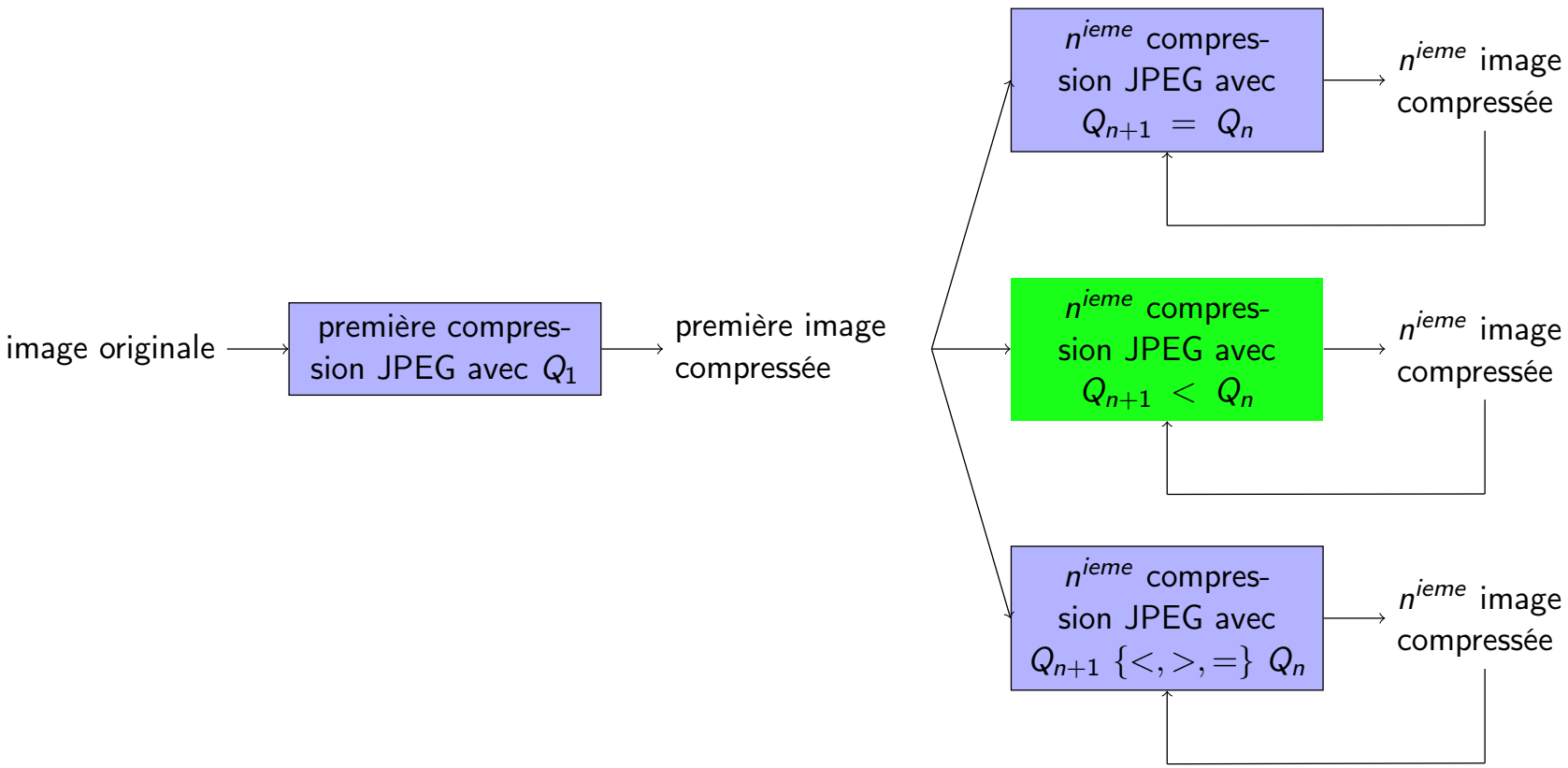
Set of Near-Duplicates

Image phylogeny tree

Deux parties importantes lors de la reconstruction de l'arbre phylogénétique :

- Correctement identifier la racine
- Estimer au mieux l'arborescence

Notre cas d'étude



Sommaire

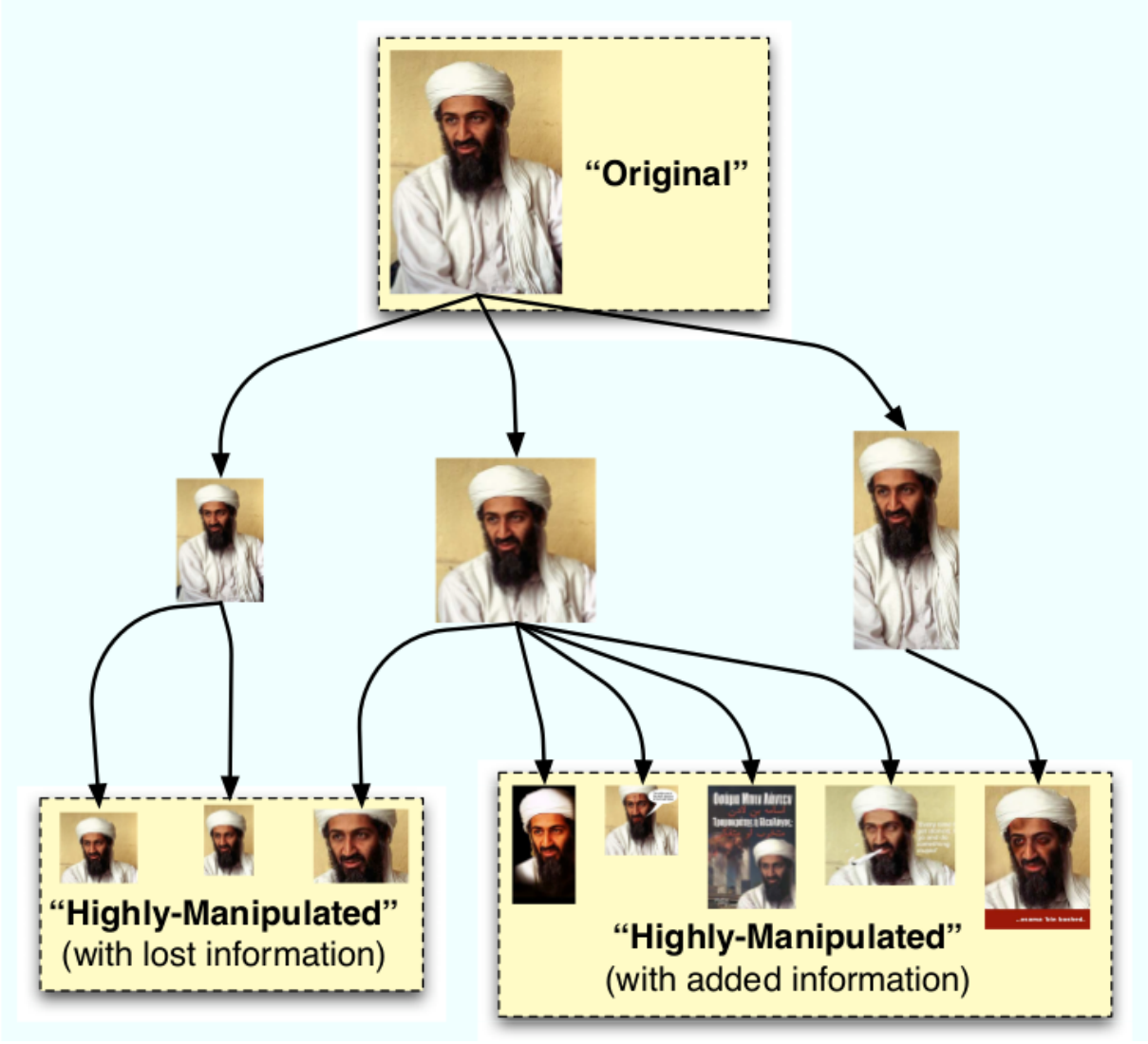
Estimation de l'arbre de phylogenie

Visual Migration Map [2]

- Les transformations sont directionnelles
- Relation parent-enfant si tous les détecteurs s'accordent sur la direction
- Simplification du graphe par sélection des plus longs chemins

2. Lyndon Kennedy et 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, p. 349–358.

Estimation de l'arbre de phylogenie



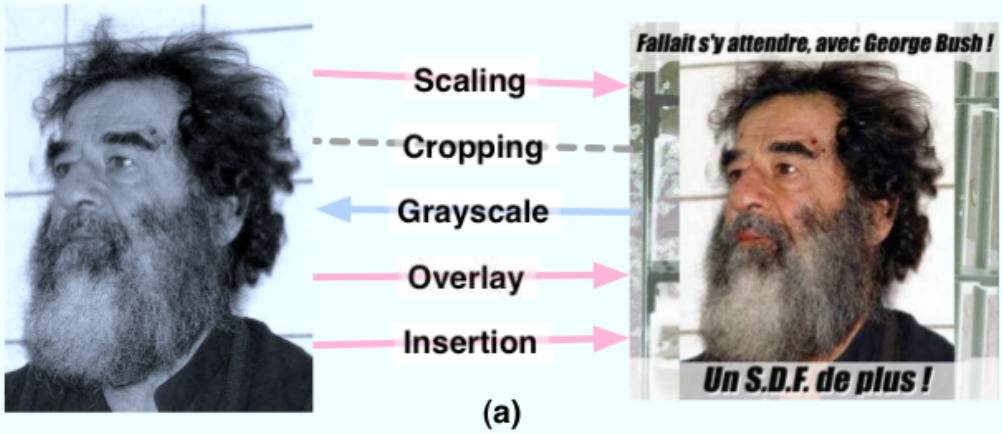
Estimation de l'arbre de phylogenie

Visual Migration Map [2]

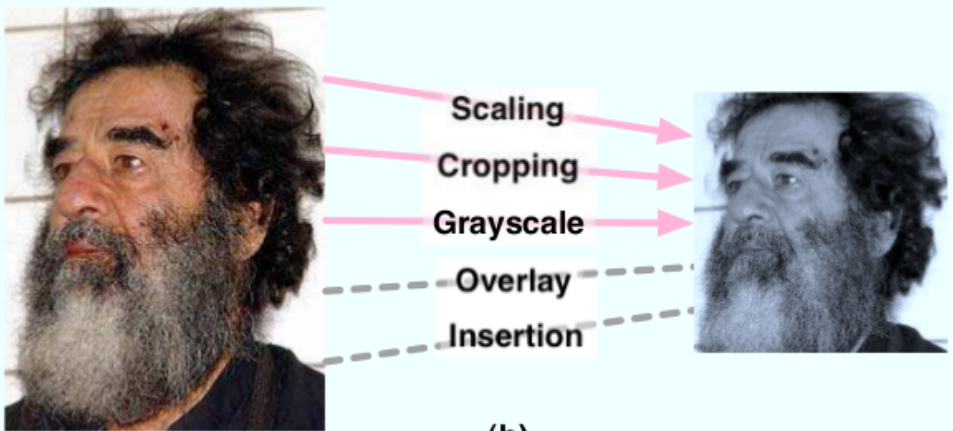
- Les transformations sont directionnelles
- Relation parent-enfant si tous les détecteurs s'accordent sur la direction
- Simplification du graphe par sélection des plus longs chemins

2. Lyndon Kennedy et 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, p. 349–358.

Estimation de l'arbre de phylogenie



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



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

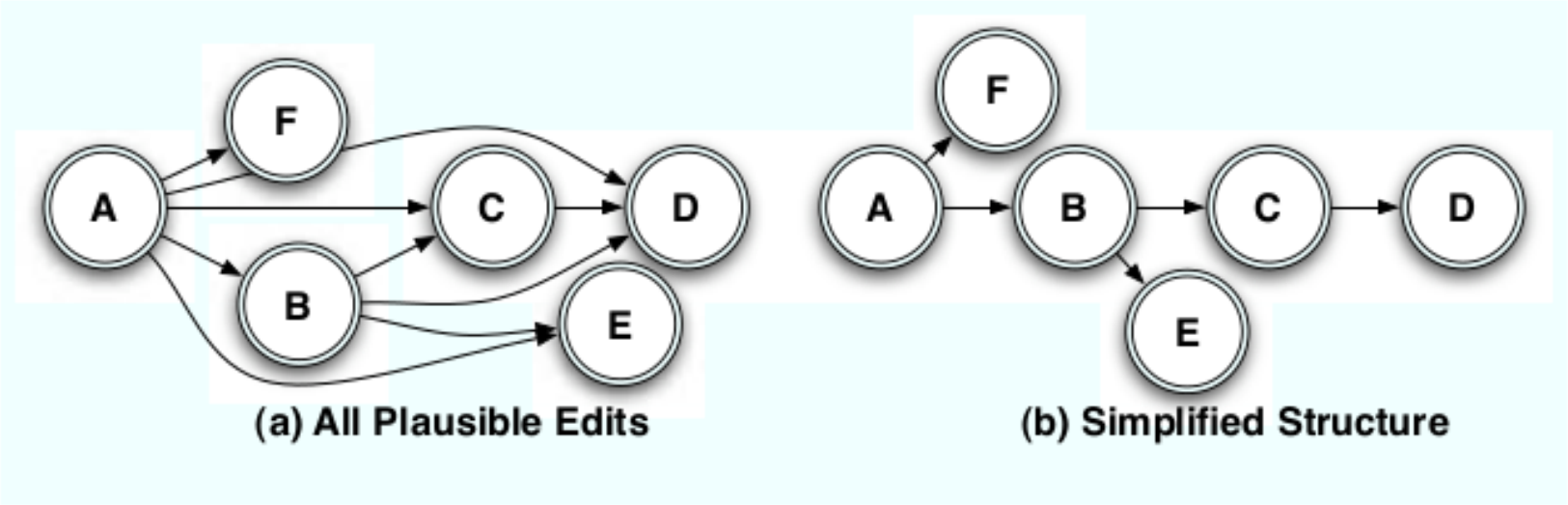
Estimation de l'arbre de phylogenie

Visual Migration Map [2]

- Les transformations sont directionnelles
- Relation parent-enfant si tous les détecteurs s'accordent sur la direction
- Simplification du graphe par sélection des plus longs chemins

2. Lyndon Kennedy et 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, p. 349–358.

Estimation de l'arbre de phylogenie



Estimation de l'arbre de phylogénie

Image phylogeny tree [3] [4]

- Calcul d'une *dissimilarity matrix*
- Calcul d'un arbre couvrant de poids min (Kruskal ou autre)

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

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

Convergence des blocs lors de compressions successives [5]

But

Compter le nombre de compressions

3 types de blocs

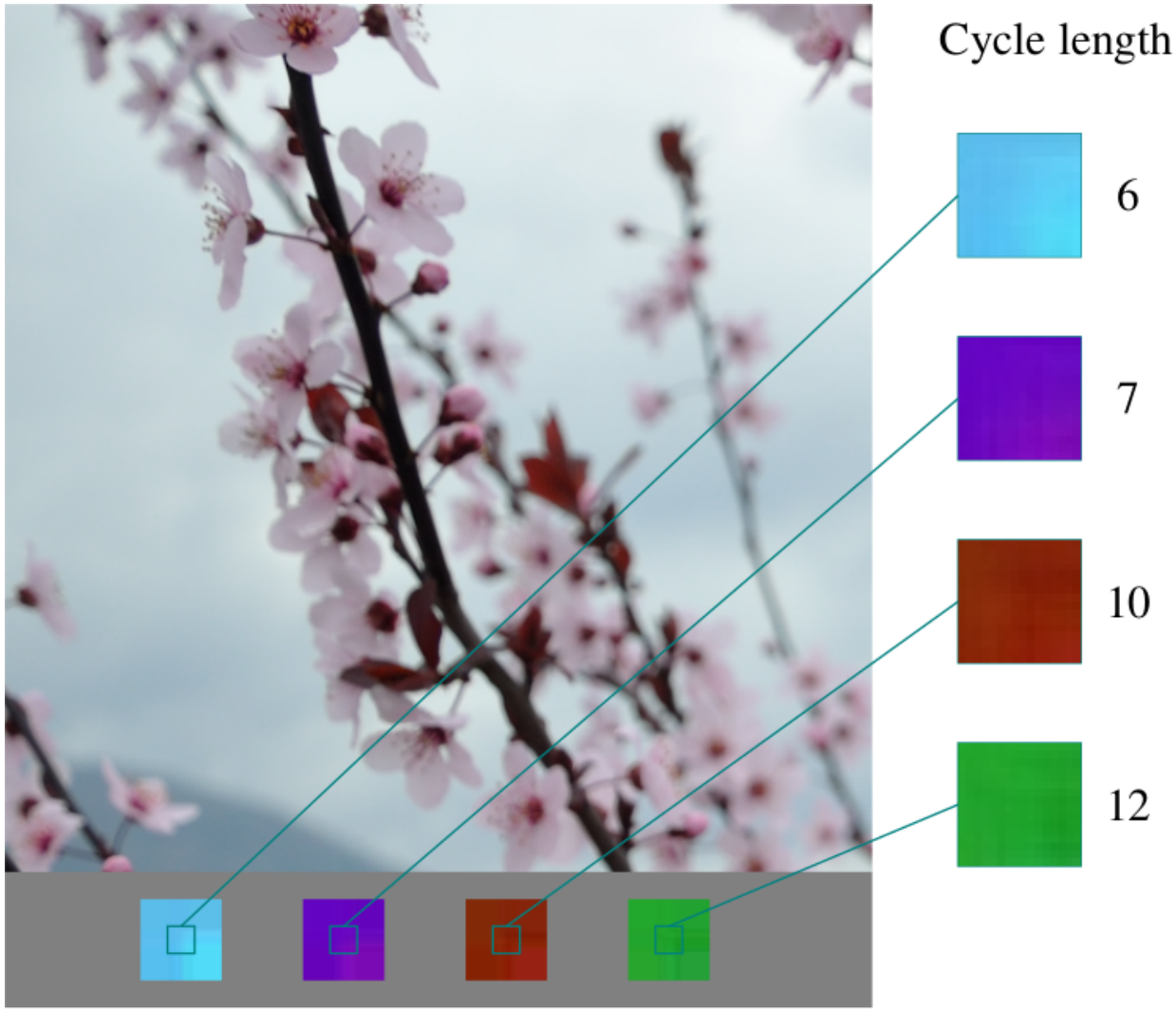
- Les blocs plats
- Les blocs stables
- Les blocs cycliques

Comment ?

plus petit commun multiple de la longueur des cycles

5. Matthias Carnein, Pascal Schöttle et Rainer Böhme. "Telltale Watermarks for Counting JPEG Compressions". In : *Proceedings of the Electronic Imaging 2016*. Publication status : Published. San Francisco, USA, 2016.

Convergence des blocs lors de compressions successives



Estimation de la matrice de quantification primaire [6]

Principe de leur méthode

Comparer l'histogramme de l'image originale et l'histogramme des images compressées avec des tables de quantification modèles puis compressées avec Q_{f_2} et enfin garder la table pour laquelle la différence entre histogramme est la plus faible

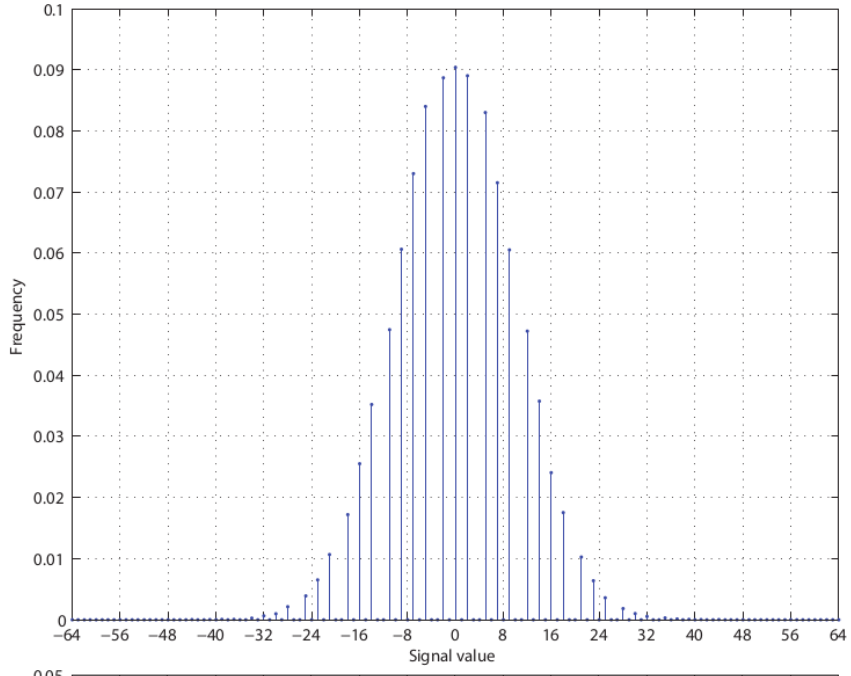
6. Jan Lukáš et Jessica Fridrich. "Estimation of primary quantization matrix in double compressed JPEG images". In : *Proc. Digital Forensic Research Workshop*. 2003, p. 5–8.

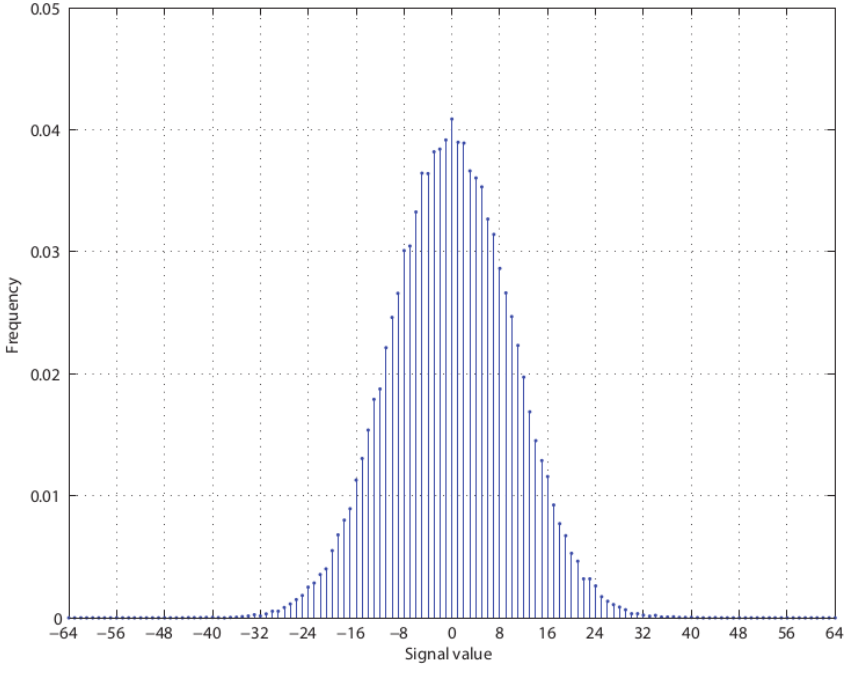
Analyse des valeurs manquantes

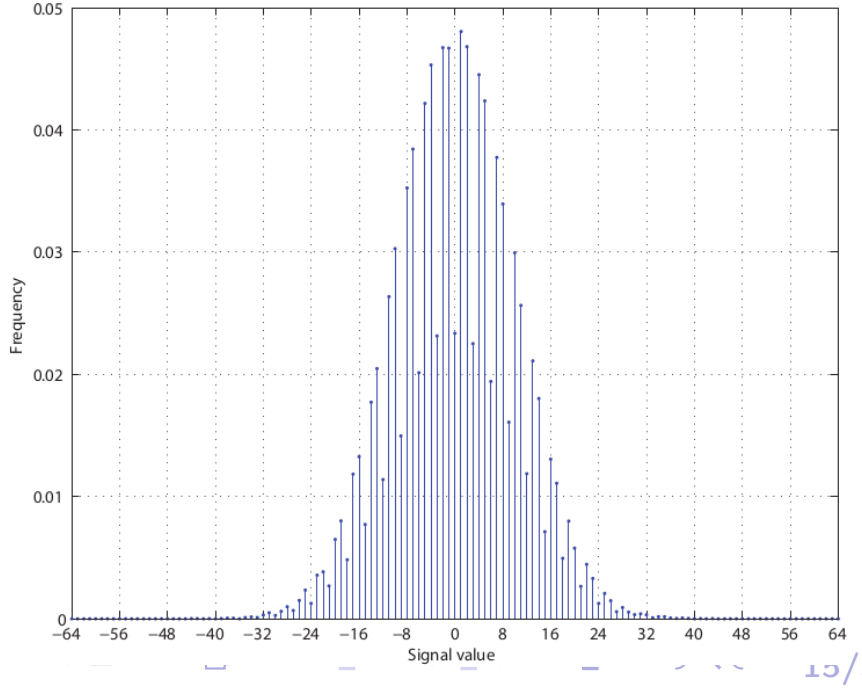
Artefacts distincts pour $Q_{f_1} > Q_{f_2}$ et $Q_{f_1} < Q_{f_2}$

Limites

- $Q_{f_1} = Q_{f_2}$
- Q_{f_1} est facteur de Q_{f_2}







Sommaire

But

Réduction d'un problème de reconstruction d'un arbre de phylogénie à un problème de **négation de parenté**

Solution

Décision binaire entre deux images : "Cette image est-elle le parent de cette autre image ?"

Notre approche

Marqueur

Caractéristique de l'image qui indique qu'une certaine opération a été effectuée et qui va se transmettre aux enfants

Notre approche

Marqueur

Caractéristique de l'image qui indique qu'une certaine opération a été effectuée et qui va se transmettre aux enfants

Fonction de négation

$f(I_m, I_n)$ est une fonction qui pour tout couple d'images (I_m, I_n) détecte à chaque fois qu'il est présent un marqueur visible dans une image et pas dans l'autre, et donc prouve qu'il n'y a pas de relation de parenté entre I_m et I_n.

Notre approche

Marqueur

Caractéristique de l'image qui indique qu'une certaine opération a été effectuée et qui va se transmettre aux enfants

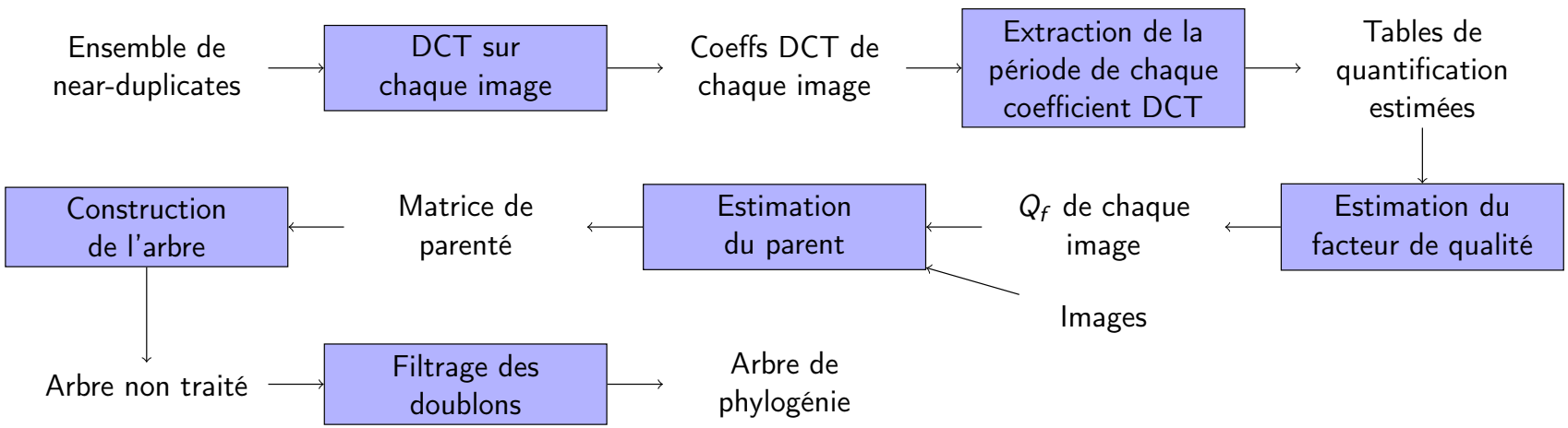
Fonction de négation

$f(I_m, I_n)$ est une fonction qui pour tout couple d'images (I_m, I_n) détecte à chaque fois qu'il est présent un marqueur visible dans une image et pas dans l'autre, et donc prouve qu'il n'y a pas de relation de parenté entre I_m et I_n.

Théorème

Pour tout couple d'images (I_m, I_n) d'un ensemble de near-duplicates, s'il n'existe pas de marqueur prouvant que I_m n'est pas le parent de I_n, alors il y a une relation parent-enfant entre I_m et I_n, $I_m \rightarrow I_n$.

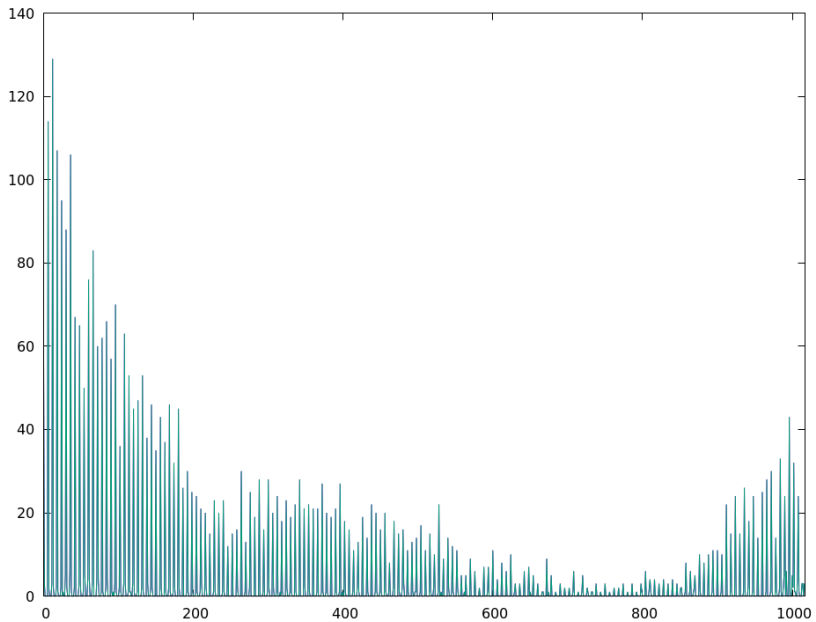
Schéma de notre approche

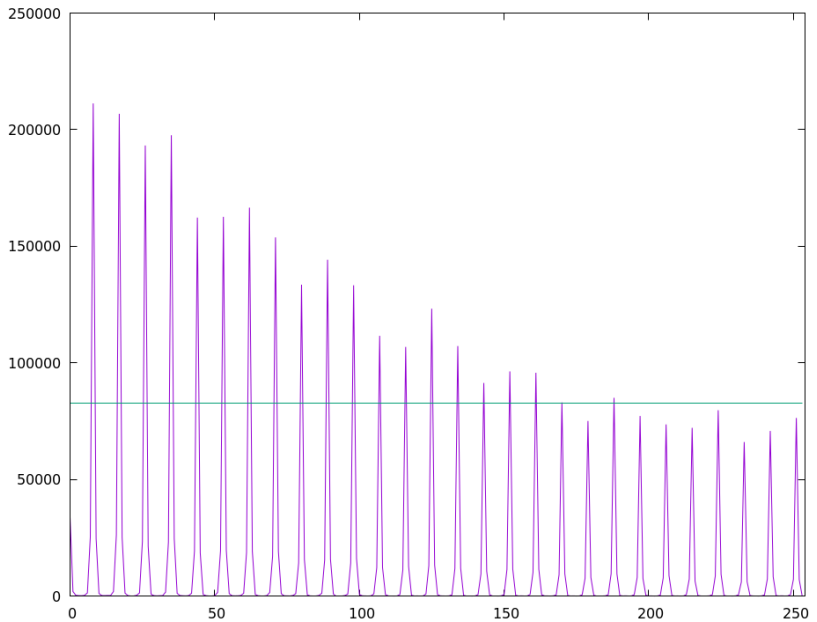


Extraction de la période

Qu'est ce que la période

Delta entre chaque pic de l'autocorrélation





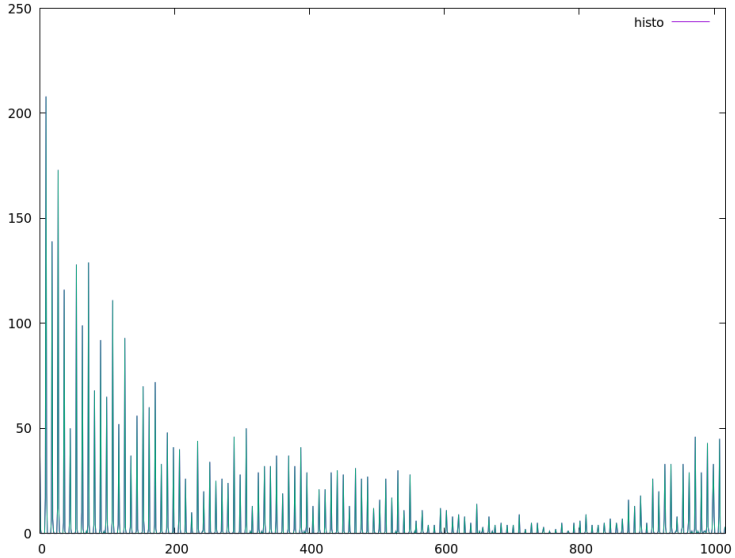
Extraction de la période

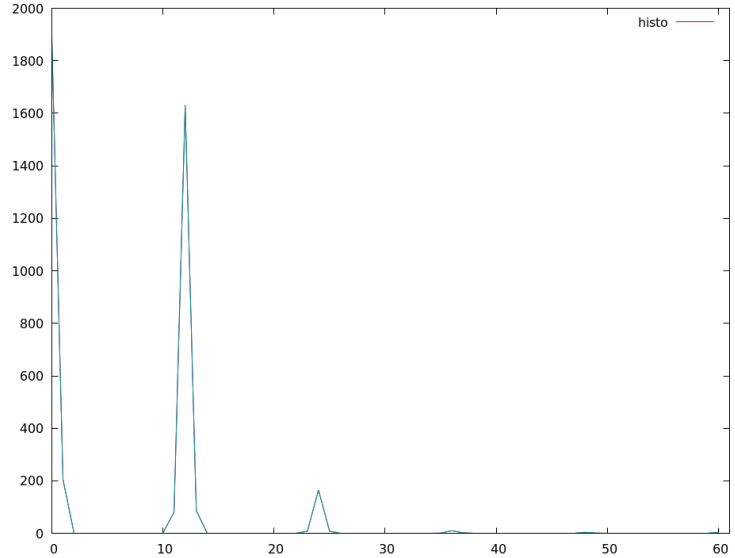
9	6	5	9	13	22	22	30
6	6	8	10	14	16	30	
8	7	9	13	20	31		
8	9	12	19	28			
10	12	-1	-1				
12	13	30					
28	31						

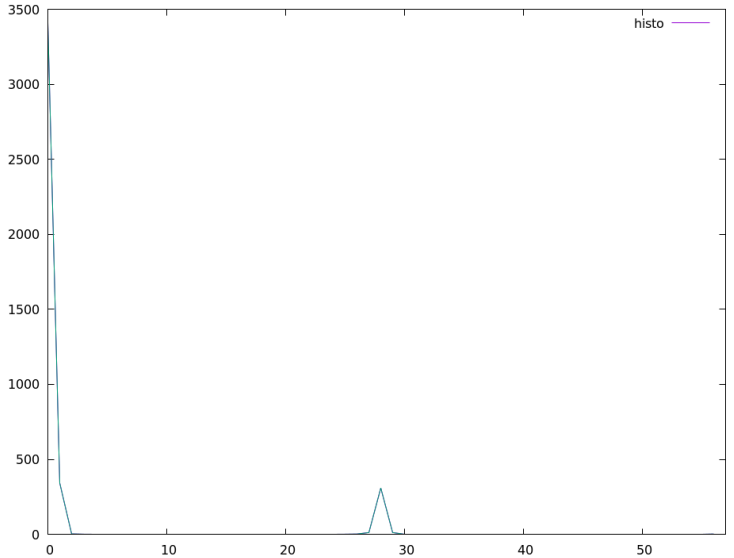
Figure – Exemple de table de quantification retournée par l'estimation de la période, $\widehat{q}(u, v)$.

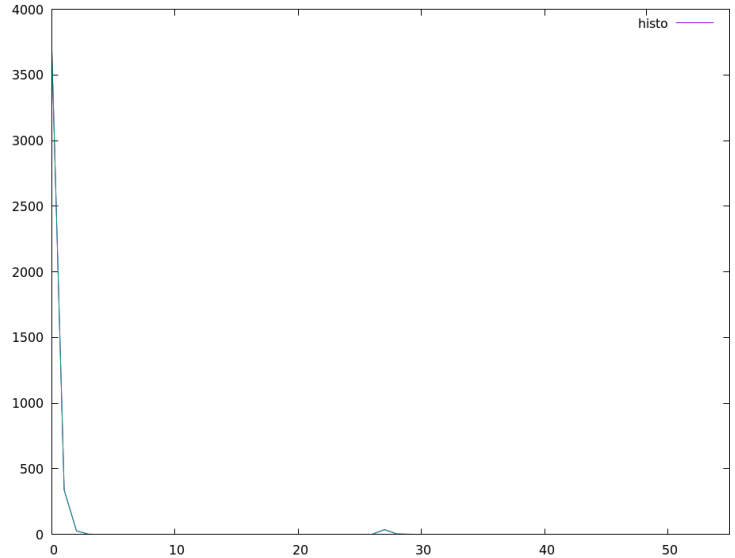
Limitation au 35 premiers coefficients

Extraction de la période









Estimation du facteur de qualité : estimation primaire

Estimation primaire

Calcul de distance entre $\widehat{q}(u, v)$ et table(i) :

$$D_{euc}(P, Q) = \sqrt{\sum_{i=1}^{d} |P_i - Q_i|^2}$$

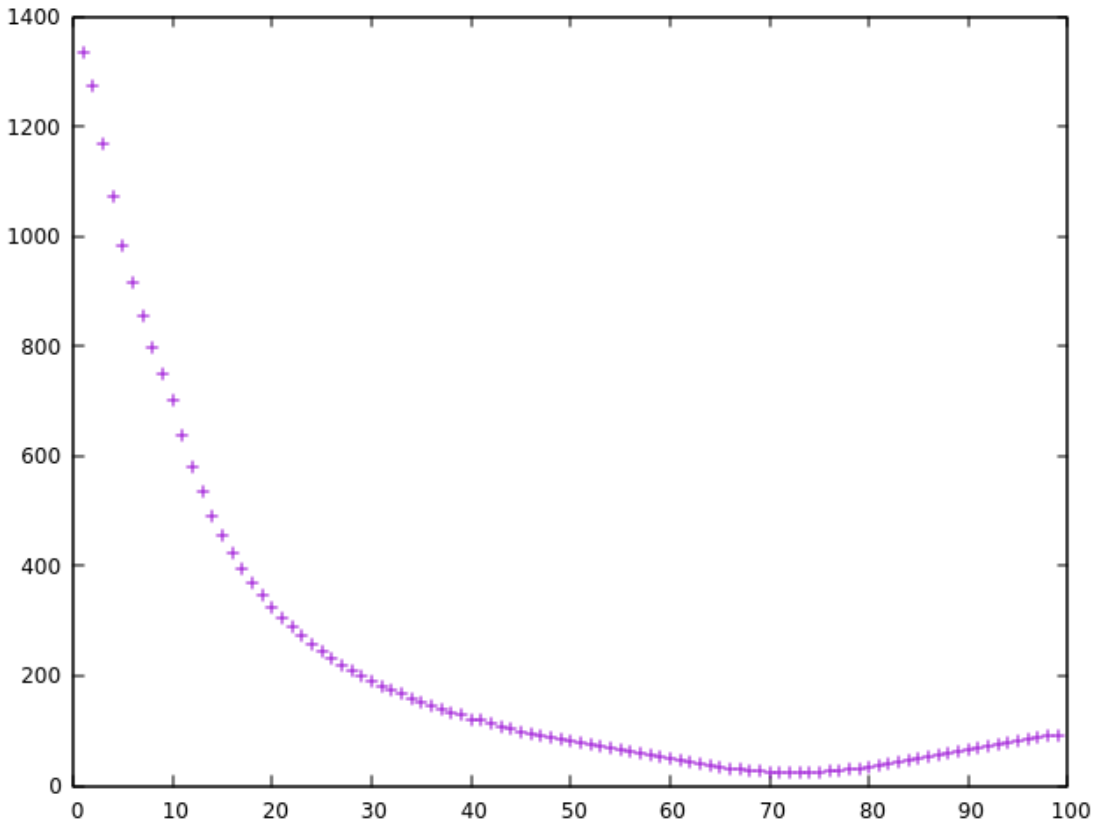
Avantage

Donne directement Q_f

Inconvénients

- Lent
- Imprécise

Estimation du facteur de qualité : estimation primaire



32	24	28	28	36	48	98	144
22	24	26	34	44	70	128	184
20	28	32	44	74	110	156	190
32	38	48	58	112	128	174	196
48	52	80	102	136	162	206	224
80	116	114	174	218	208	242	200
102	120	138	160	206	226	240	206
122	110	112	124	154	184	202	198

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

Estimation du facteur de qualité : estimation secondaire

Estimation secondaire

Utilisation des formules

Avantages

- Rapide
- Précise

Inconvénients

Q_f est nécessaire pour calculer Q_f

$$Si \quad Q_f < 50 \quad Q_s = 5000/Q_f \quad sinon \quad Q_s = 200 - (Q_f \times 2) \quad (1)$$

$$q(u,v) = \frac{(base(u,v) \times Q_s) - 50}{100} \quad avec \quad 1 \leq q(u,v) \leq 255 \quad (2)$$

Estimation des ancêtres

La compression est déterministe

La compression n'est pas transitive

Le parent est à une compression de distance de ses enfants

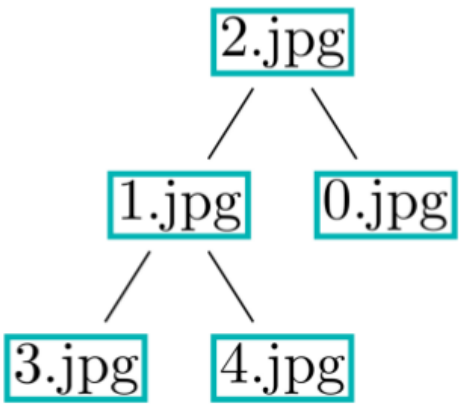
Suppression des images ne pouvant pas être un ancêtre

Décision binaire

Reconstruction de l'arbre

Matrice binaire de taille $n \times n$

Construction de l'arbre à partir de la matrice



(a) Arbre de phylogénie

-	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) Matrice de parenté

Points clés de notre approche

Réduction d'un problème de reconstruction d'un arbre de phylogénie à un problème de **négation de parenté**

Facilement extensible

Sommaire

Arbres complets

Une très bonne estimation du Q_f

Des métriques proches de 100%

La précision baisse quand la taille de l'arbre augmente

Arbres complets

Métrique \ Dataset	15 images	25 images	50 images
Erreur moyenne d'estimation de Q_f	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

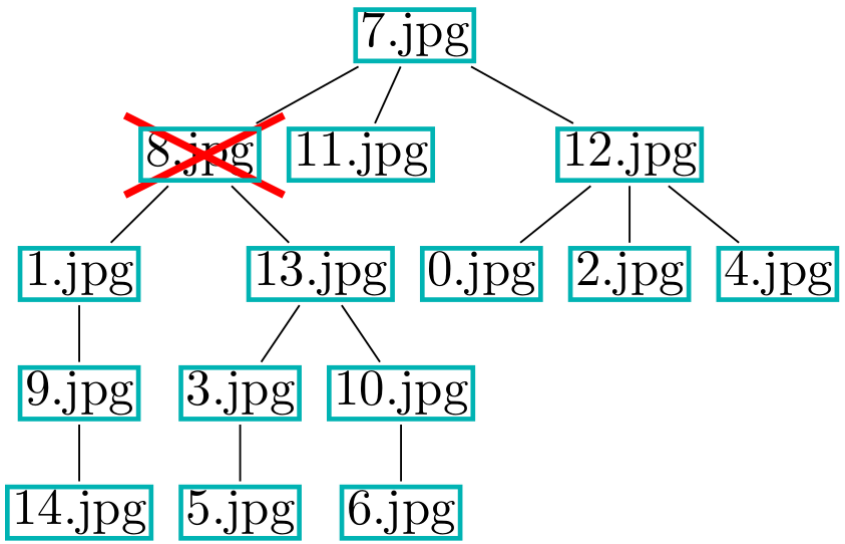
Arbres avec une image manquante

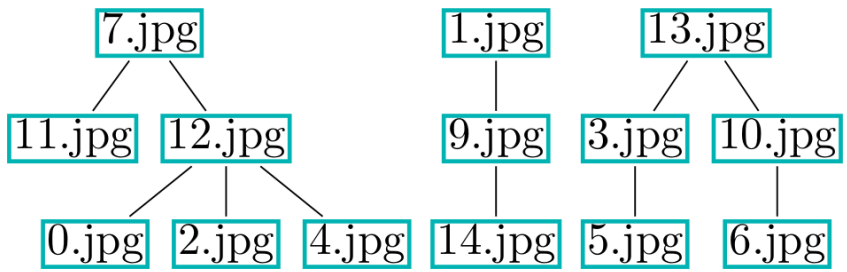
Mauvaise estimation de la racine

Bonne estimation du reste de l'arbre

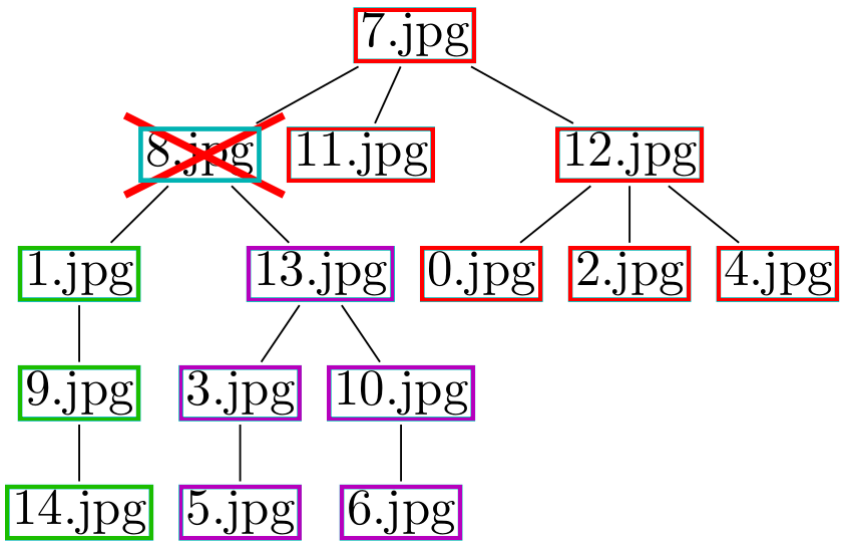
Notre méthode ne détecte que le parent

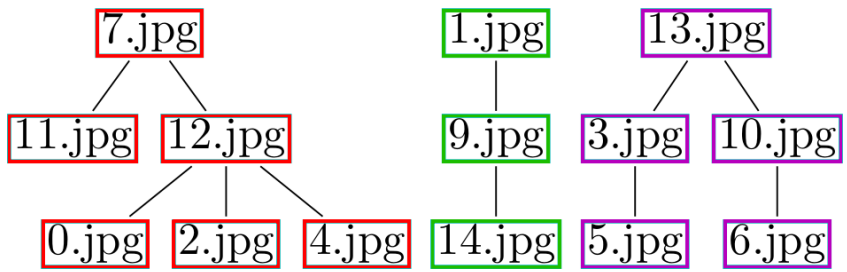
Arbres avec une image manquante





Arbres avec une image manquante





Arbre avec des images en couleur

- Aucune adaptation de notre implémentation
- Utilisation du canal de luminance
- De très bons résultats

Arbre avec des images en couleurs

Métrique \ Dataset	15 images	25 images	50 images
Erreur moyenne d'estimation de Q_f	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

Sommaire

Conclusion - perspectives

- Une méthode prometteuse
- Trouver d'autres marqueurs
- Traiter tous les cas de la compression JPEG
- Ne pas se limiter à la compression

Conclusion - perspectives

Des questions ?