# La phylogénie des images dans les réseaux sociaux


## Étude bibliographique

Master *Sciences et Technologies*,
Mention *Informatique*,
Parcours IMAGINA

**Auteur**
Noé Le Philippe

**Superviseurs**
William Puech

**Lieu de stage**
Équipe ICAR - LIRMM UM5506 - CNRS, Université de Montpellier

version 1 9 mars 2016

**Résumé**

Ce stage de master.

---

**Abstract**

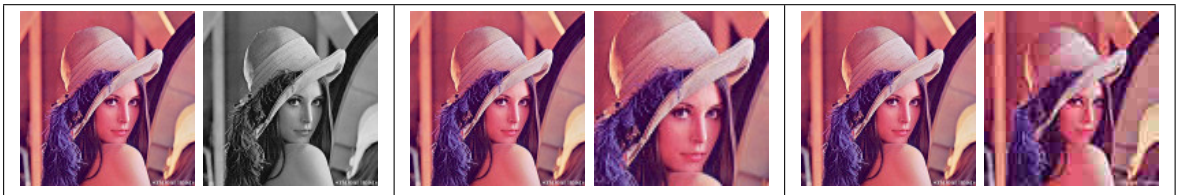
This master thesis.

# Table des matières

# 1

# Introduction

La phylogénie, en sciences naturelles, est définie[20] comme l'étude des relations de parenté entre êtres vivants. Et c'est exactement de cela qu'il s'agit dans le cas des images, l'étude des relations de parenté entre images.
À l'ère des réseaux sociaux, il n'a jamais été aussi simple de partager des idées et du contenu. À chaque partage cependant, l'information peut être amenée à être modifiée. Les images, puisque c'est là notre sujet d'étude, peuvent avoir subi un certain nombre de transformations et modifications avant de nous parvenir. C'est dans ce contexte que nous allons intervenir, et tenter de reconstituer la phylogénie de l'image. Il peut être difficile de différentier cette image de l'originale, et de savoir laquelle est l'originale, mais c'est pourtant crucial dans un monde où l'information peut être falsifiée par tout le monde et extrêmement facilement. Les applications sont variées, et ne se cantonnent pas à la détection et la discrimination d'images altérées, on peut également se servir de la phylogénie de l'image pour optimiser de l'espace de stockage en ne gardant que l'originale ou encore suivre la diffusion et l'évolution des idées sur les réseaux sociaux.



couleur → noir et blanc | image entière → crop | image → image compressée

FIGURE 1.1 : Exemples de near-duplicates

### Near-duplicate images (NDI)

Nous travaillons sur un ensemble d'images, toutes similaires, et au milieu de cette jungle d'images, nous devons décider quelle image est le parent de quelle autre, ou autrement dit, quelles images sont des **near-duplicates**. Joly et al. [12] définit la notion de near-duplicate comme suit : $I_1 = T(I), T \in \mathcal{T}$ où $I$ est l'image parent, $I_1$ est l'image enfant et $\mathcal{T}$ est un ensemble de transformations autorisées, $I_1$ et $I$ sont alors des NDI. Dans le cas général, $\mathcal{T} = \{$*resampling, cropping, affine warping, color changing, lossy compression*$\}$, Fig. 1 montre un exemple de near-duplicates, dans le cadre de ce stage cependant, $\mathcal{T} = \{$*lossy compression*$\}$. Notons l'utilisations du terme *transformations autorisées*. Ce terme est double, d'une part, il place une limite arbitraire dans la force de la transformation, par exemple une image cropée à plus de 10% pourra ne pas pas être considérée comme un near-duplicate, et d'autre part, il permet de restreindre l'espace des transformations possibles. Ainsi, dans le cadre du stage, seules les images de la troisième case de Fig. 1 seront des NDI. Ces transformations peuvent évidemment se composer, et une image enfant peut être le résultat de plusieurs transformations.

### Arbre phylogénétique (Image Phylogeny Tree - IPT)

C'est l'arbre représentant les relations de parenté entre les différentes images. Il sera extrait d'un ensemble de NDI, c'est l'objectif final de l'application. Un exemple est disponible Fig. 1.2. Le passage d'une génération à l'autre, autrement dit d'un noeud à son fils se fait à travers la transformation $I_1 = T(I)$, ainsi, une image et son parent sont des NDI, alors qu'une images et ses soeurs ne le sont pas (Fig 1.3).

La reconstruction de l'arbre se concentre autour de deux problèmes principaux. Le premier est l'identification de la racine, et le second et l'estimation du reste de l'arbre. Il est en effet critique d'identifier correctement la racine. Prenons par exemple un des cas d'utilisation de l'IPT, la détection d'altération d'images. L'idée est que pour un ensemble d'images, plus on est proche de la racine, moins l'image a subi de transformations, et donc moins elle est altérée, avec dans le meilleur des cas, la racine en image originale. On voit bien que si on identifie mal la racine, on déduira, à tort, qu'une image n'a pas été altérée. Il n'est pas toujours garanti que la totalité des images de l'arbre original soit présente, de plus, certaines transformations peuvent être mineures, et difficile à détecter, c'est donc bien une estimation de l'arbre original qui sera faite.

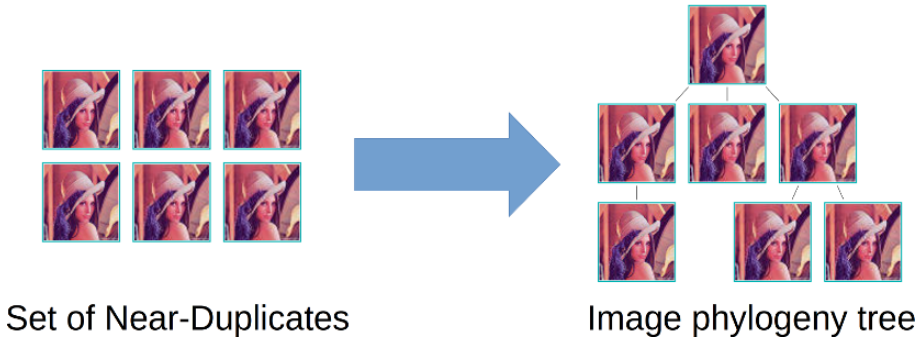



FIGURE 1.2 : Passage d'un ensemble de NDI à un arbre phylogénétique

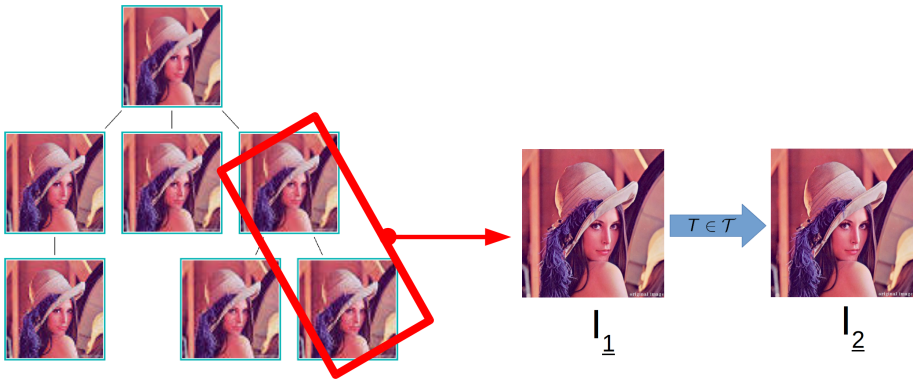



FIGURE 1.3 : Passage d'une image parent à l'image enfant

### Pourquoi se restreindre à la compression ?

Comme mentionné précédemment, nous ne considérons que le cas de la compression avec perte en laissant de coté les autres transformations. Nous avons décidé de ne nous concentrer que sur une seule transformation pour avoir la possibilité de la traiter en détail et en profondeur dans le cadre du stage et ne pas devoir survoler sans approfondir toutes les transformations. Le choix de la compression est assez naturel, cela n'a, à notre connaissance, pas encore été traité dans le cadre le la phylogénie, et c'est un domaine largement étudié en forensic.

# 2

# État de l'art

## 2.1 État de l'art - Étude de l'arbre phylogénétique

C'est un sujet qui, ces dernières années, n'a retenu l'attention que de deux équipes de recherche, nous allons ici présenter leurs travaux.

### La Visual Migration Map (VMM)

C'est à notre connaissance, le premier article concernant vraiment notre sujet. Kennedy et al. [13] proposent une approche permettant d'automatiquement détecter la manière dont une image a été éditée ou manipulée, et d'en extraire des relations parent-enfant entre les images. Il vont construire à partir le l'estimation de ces transformations une Visual Migration Map (VMM) (voir Fig. 2.1) qui est en fait notre arbre de phylogénie.

Ils partent du principe que les transformations sont directionnelles, c'est à dire que l'on de ne peut passer que d'une image moins transformée à une image plus transformée. Ainsi, ils vont tenter d'estimer la direction de chaque transformation entre deux images $I_1$ et $I_2$ (sachant que $\mathcal{T} = \{$*scaling, cropping, grayscale, overlay, insertion*$\}$). Trois scénarios sont alors possibles : si toutes des transformations sont dans le même sens, l'image fille est alors celle vers qui pointent les transformations, si les transformations sont dans des sens contraires, les images sont surement des soeurs, elles n'ont ent tous cas pas de relation parent-enfant, et enfin si aucune transformation n'a été détectée, c'est que soit les images sont identiques, soit elles ne sont pas des near-duplicates. On peut en voir un exemple Fig. 2.1.

Un graphe va ensuite être construit à partir des couples d'images pour lesquels une relation parent-enfant a été détectée. À noter qu'une relation parent-enfant ne veut pas forcément dire que c'est le parent direct mais plutôt un ancêtre. Ainsi, un noeud du graphe (une image) peut avoir plusieurs noeuds parents, pour finalement obtenir l'arbre désiré, seuls les chemins les plus longs sont conservés, comme on peut le voir Fig. 2.1.

### Image Phylogeny Tree (IPT)

## 2.2 État de l'art - Analyse des recompression JPEG

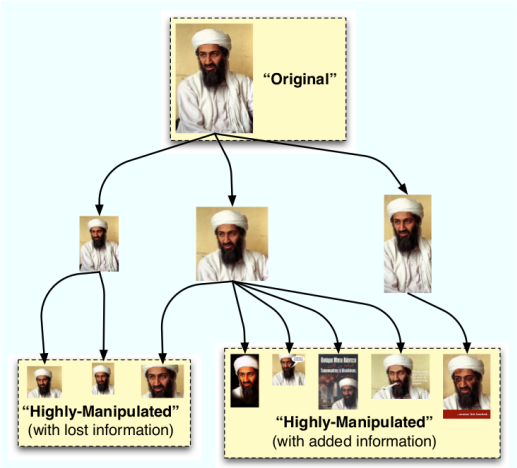



FIGURE 2.1 : Exemple de VMM, issu de [13]

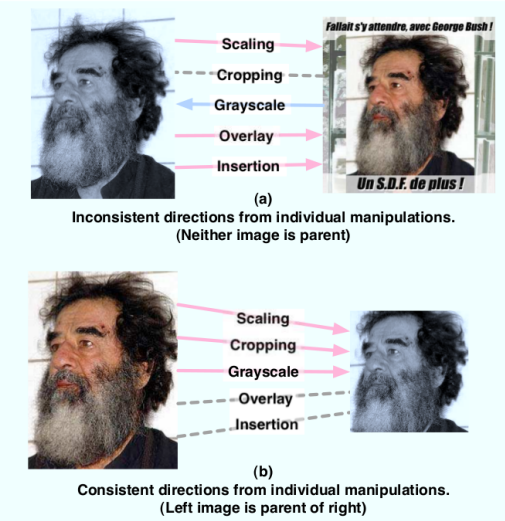



FIGURE 2.2 : Exemple de direction des transformations, issu de [13]

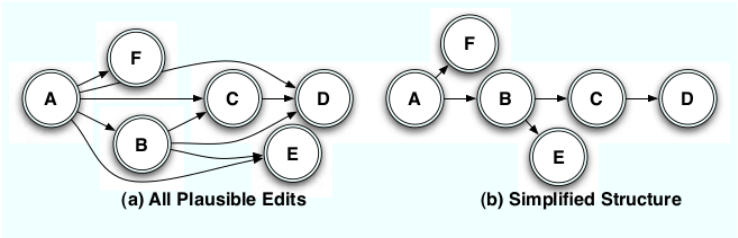



FIGURE 2.3 : Exemple de simplification de graphe, issu de [13]

# 3

# Notre approche

# Conclusion

# Bibliographie


[1] Tiziano Bianchi et Alessandro Piva. “Detection of non-aligned double JPEG compression with estimation of primary compression parameters”. In : *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE. 2011, p. 1929–1932.

[2] Tiziano Bianchi et Alessandro Piva. “Detection of nonaligned double JPEG compression based on integer periodicity maps”. In : *Information Forensics and Security, IEEE Transactions on* 7.2 (2012), p. 842–848.

[3] Tiziano Bianchi et Alessandro Piva. “Reverse engineering of double JPEG compression in the presence of image resizing”. In : *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*. IEEE. 2012, p. 127–132.

[4] Yi-Lei Chen et Chiou-Ting Hsu. “Detecting recompression of JPEG images via periodicity analysis of compression artifacts for tampering detection”. In : *Information Forensics and Security, IEEE Transactions on* 6.2 (2011), p. 396–406.

[5] Sung-Hyuk Cha. “Comprehensive survey on distance/similarity measures between probability density functions”. In : *City* 1.2 (2007), p. 1.

[6] Zanoni Dias, Siome Goldenstein et Anderson Rocha. “Exploring heuristic and optimum branching algorithms for image phylogeny”. In : *Journal of Visual Communication and Image Representation* 24.7 (2013), p. 1124–1134.

[7] Zanoni Dias, Siome Goldenstein et Anderson Rocha. “Large-scale image phylogeny : Tracing image ancestral relationships”. In : *MultiMedia, IEEE* 20.3 (2013), p. 58–70.

[8] 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.

[9] 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.

[10] Hany Farid. “Exposing digital forgeries from JPEG ghosts”. In : *Information Forensics and Security, IEEE Transactions on* 4.1 (2009), p. 154–160.

[11] Fangjun Huang, Jiwu Huang et Yun Qing Shi. "Detecting double JPEG compression with the same quantization matrix". In : *Information Forensics and Security, IEEE Transactions on* 5.4 (2010), p. 848–856.

[12] 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.

[13] 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.

[14] ShiYue Lai et Rainer Bohme. "Block convergence in repeated transform coding : JPEG-100 forensics, carbon dating, and tamper detection". In : *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, p. 3028–3032.

[15] Edmund Y Lam et Joseph W Goodman. "A mathematical analysis of the DCT coefficient distributions for images". In : *Image Processing, IEEE Transactions on* 9.10 (2000), p. 1661–1666.

[16] Bin Li et al. "JPEG noises beyond the first compression cycle". In : *arXiv preprint arXiv :1405.7571* (2014).

[17] Marina A Oikawa et al. "Manifold learning and spectral clustering for image phylogeny forests". In : *Information Forensics and Security, IEEE Transactions on* 11.1 (2016), p. 5–18.

[18] A Oliveira et al. "Multiple parenting identification in image phylogeny". In : *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE. 2014, p. 5347–5351.

[19] Alberto A de Oliveira et al. "Multiple parenting phylogeny relationships in digital images". In : *Information Forensics and Security, IEEE Transactions on* 11.2 (2016), p. 328–343.

[20] *Phylogénie*. URL : `https://fr.wikipedia.org/wiki/Phylogen%C3%A8se`.

[21] Giovanni Puglisi et al. "First JPEG quantization matrix estimation based on histogram analysis". In : *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE. 2013, p. 4502–4506.

[22] Jianquan Yang et al. "An Effective Method for Detecting Double JPEG Compression With the Same Quantization Matrix". In : *Information Forensics and Security, IEEE Transactions on* 9.11 (2014), p. 1933–1942.