Phylogenetic Analysis of Near-Duplicate Images using Processing Age Metrics

S. Milani[1], M. Fontana[1], P. Bestagini[2], S. Tubaro[2]

1 :Department of Information Engineering, University of Padova, Italy
2 :Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

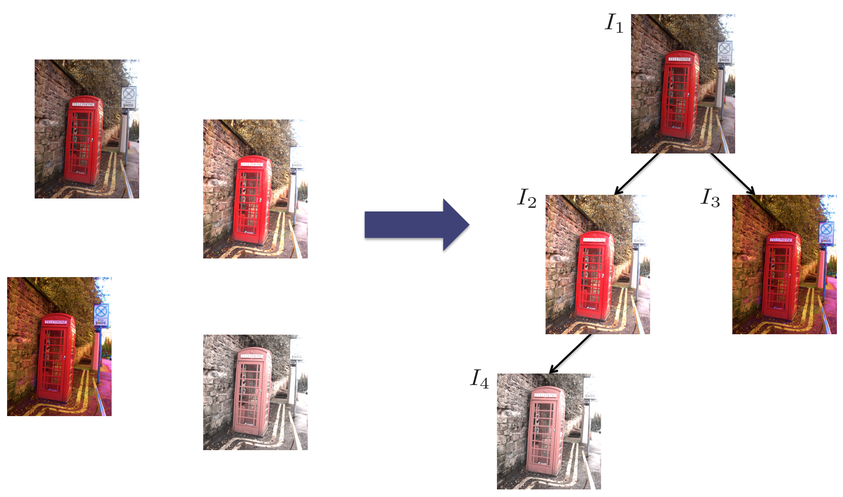
1-Introduction

Recent researches on image forensics have led to the design of algorithms to study the phylogenetic relationship between near-duplicate (ND) images.
In this work we propose a set of features that blindly model the processing age of an image, i.e., how much an image has been edited in its lifetime. By exploiting these features, it is possible to improve the performance of phylogenetic relationship reconstruction algorithms by increasing their accuracy and reducing computational complexity.

2-Problem formulation

Problem
Given a set of ND images, reconstruct the image phylogeny tree (IPT).

Image Phylogeny Tree (IPT)
The IPT is an acyclic directed graph representing parental relationships.



3-IPT reconstruction

Rationale
If I_c (child image) has been generated from I_p (parent image) it is possible to map I_p to I_c but not the vice versa.

Pipeline

- Image registration: search for the best estimate $I_{p\to c}$ of I_c obtained from I_p.
- Dissimilarity computation: evaluate the goodness of the estimated $I_{p\to c}$ computing image dissimilarity $d_{p,c} = \mathcal{L}(I_c, I_{p\to c})$, where $\mathcal{L}$ is a distance metric.
- IPT reconstruction: dissimilarity matrix $D = [d_{p,c}]$ can be interpreted as a fully connected graph. Optimum Branching (OB) algorithm [1] can be used to generate an IPT from D.

4-Processing age

Motivation

- $I_{p\to c}$ computation is time consuming and performed for every image pairs

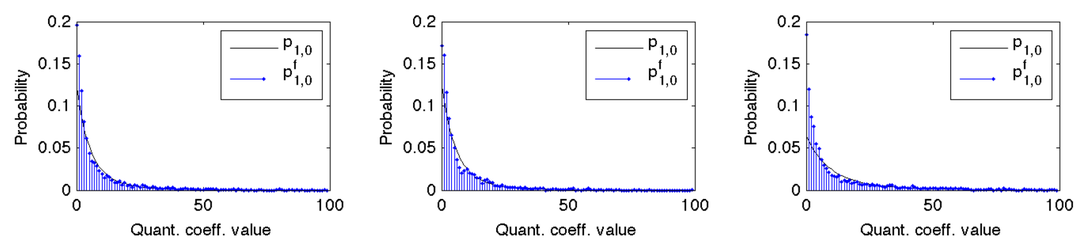
Proposed solution

- Rank images according to their processing age.
- Compute $I_{p\to c}$ only for images in young-old relationship.

5-Processing age metrics

Idea
Image processing operations scramble DCT coefficients statistics.
The divergence between computed statistics and a fitting model is proportional to the processing age.

- probability mass function of (i,j) DCT coefficient: $p_{i,j}(c)$
- fitting model: $p^f_{i,j}(c) = \Gamma e^{-\pi(c)}$
- processing age: PA-$Xc = \frac{D_X(p_{1,0}(c)||p^f_{1,0}(c)) + D_X(p^f_{1,0}(c)||p_{1,0}(c))}{2}$



The same applies to DCT coefficients first digit (e.g., Benford's law)

- first digit: $m = \mathrm{FD}_M(c) = \left\lfloor \frac{|c|}{M^{\lfloor \log_M |c| \rfloor}} \right\rfloor$.

Divergences

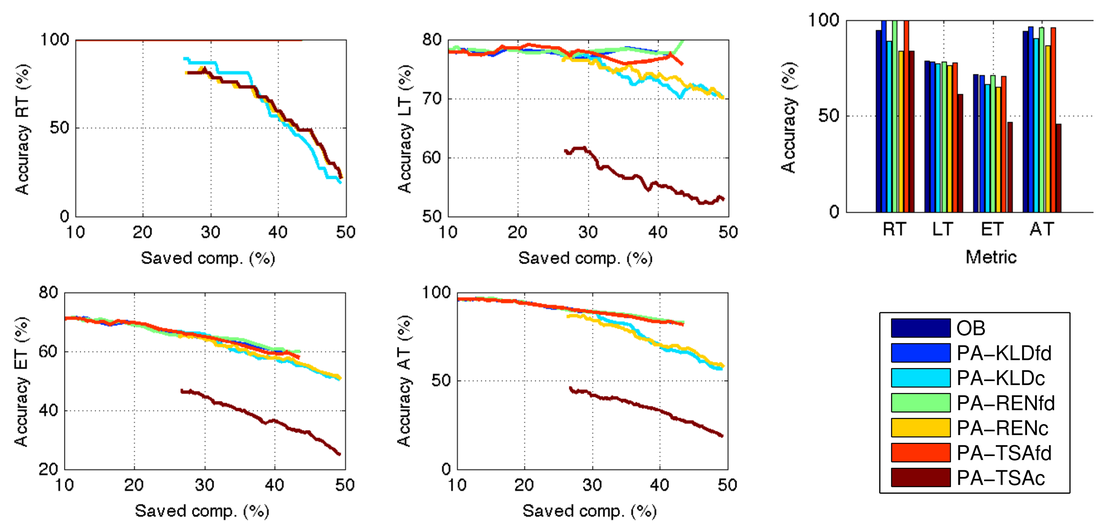
- Kullback-Leibler: $D_{KL}(p||p^f) = \sum_c p(c) \log_2 \frac{p(c)}{p^f(c)}$
- Renyi: $D^\alpha_R(p||p^f) = \frac{1}{\alpha - 1} \log_2 \left(\sum_c p(c)^\alpha p^f(c)^{1-\alpha} \right)$
- Tsallis: $D^\alpha_T(p||p^f) = \frac{1}{1-\alpha} \left(1 - \sum_c (p(c)^\alpha p^f(c)^{1-\alpha}) \right)$

6-Experiments

Dataset

- 50 trees of 10 and 30 nodes for a total number of $50 \times (10+30) = 2000$ images.
- Up to 4 transformations for each node (i.e., resampling, cropping, rotation, compression).

Results



- RT: Correctly identified roots
- LT: Correctly identified leaves
- ET: Correctly identified edges
- AT: Correctly identified ancestors

References

[1] D. Zanoni, G. Siome, R. Anderson, "Exploring heuristic and optimum branching algorithms for image phylogeny", Journal of Visual Communication and Image Representation (JVCIR), 2013.