Análise Forense de Documentos Digitais

Prof. Dr. Anderson Rocha

anderson.rocha@ic.unicamp.br

http://www.ic.unicamp.br/~rocha

Reasoning for Complex Data (RECOD) Lab.
Institute of Computing, Unicamp

Av. Albert Einstein, 1251 – Cidade Universitária
CEP 13083-970 • Campinas/SP – Brasil

Summary

- Phylogeny Problem
- Image Phylogeny
- Video Phylogeny
- Current Challenges

Phylogeny of Media

Parts of this class were presented in talk in Feb. 2nd, 2012 at Google, US,
by S. Goldenstein

Joint work – S. Goldenstein, Z. Dias and A. Rocha

What is this talk about?

What is this talk about?

- An important problem that has been mostly overlooked by the community.

What is this talk about?

- An important problem that has been mostly overlooked by the community.
- Has immediate applications in many areas.

What is this talk about?

- An important problem that has been mostly overlooked by the community.
- Has immediate applications in many areas.
- It is hard to solve.

What is this talk about?

- An important problem that has been mostly overlooked by the community.
- Has immediate applications in many areas.
- It is hard to solve.
- There is room for elegant math and different solutions.

What is this talk about?

- An important problem that has been mostly overlooked by the community.
- Has immediate applications in many areas.
- It is hard to solve.
- There is room for elegant math and different solutions.

Image Phylogeny by Minimal Spanning Trees
Z Dias, and A Rocha and S Goldenstein
IEEE Transactions of Information Forensics and Security, April 2012.

Video Phylogeny: Recovering Near-Duplicate Video Relationships
Z Dias, A Rocha, and S Goldenstein
IEEE Workshop on Information Forensics and Security (WIFS) 2011

How it started

In 2009, the current Brazilian president was the president's chief of staff, and the government pre-candidate for the 2010 presidential election.

Folha de SP, a major Brazilian newspaper (think New York Times) ran an interview and article about her. They printed a "scan of her criminal records" as political activist in the military dictatorship period (1964-1985), suggesting it as a record she engaged in violent armed activities (which she denies to this day).

How it started

UM JORNAL A SERVIÇO DO BRASIL ★ ★ ★ WWW.FOLHA.COM.BR

FOLHA DE S.PAULO

DIRETOR DE REDAÇÃO: OTAVIO FRIAS FILHO

DOMINGO, 5 DE ABRIL DE 2009
ANO 89 ★ Nº 29.222

EDIÇÃO SÃO PAULO, CONCLUÍDA ÀS 22H01 ★ R$ 4,00

história

Ficha de Dilma Rousseff no Dops

Grupo de **Dilma** planejou sequestro de **Delfim Netto**

FERNANDA ODILA
DA SUCURSAL DE BRASÍLIA

Antonio Spinoza, ex-colega da ministra (Casa Civil), diz que o grupo armado que dirigiram teve como alvo o titular da Fazenda em 1969.

A ação chegou a ter data e local definidos. Um mapa da emboscada consta de processo no STM. Dilma, hoje aliada de Delfim, negou de forma "peremptória". "Ele fantasiou. Não me lembro disso." Pág. A8

Brasil gasta com 'spread' 2,5 vezes o orçamento da Saúde

Estudo calcula que pessoas físicas e jurídicas pagaram R$ 134,5 bi em 2008

Em 2008, os brasileiros gastaram R$ 134,5 bilhões em "spread" bancário – diferença entre a taxa paga pelo banco e a que é aplicada em empréstimos a consumidores. O valor é duas vezes e meia o orçamento do Ministério da Saúde no período.

Segundo estudo feito pela Fecomércio-SP (Federação do Comércio do Estado de SP), as pessoas físicas foram responsáveis por R$ 85,4 bilhões do total, e as empresas, por R$ 49,1 bilhões. O "spread" bancário no Brasil é o mais alto do mundo.

No cálculo do "spread" estão impostos, risco de inadimplência, custos e lucro.

Considerando empréstimo pessoal de R$ 1.000 a ser quitado em um ano, dos R$ 604 que o cliente de banco pagava de juros em 2008, R$ 475 equivaliam ao "spread".

O governo Luiz Inácio Lula da Silva criou um grupo que reúne técnicos do Ministério da Fazenda e do Banco Central para estudar maneiras de reduzir a margem bancária. Para especialistas, a diminuição vai demandar tempo. Págs. B1 e B3

Otan anuncia plano de enviar 5 mil militares ao Afeganistão

MARCELO NINIO
ENVIADO ESPECIAL A ESTRASBURGO

How it started

UM JORNAL A SERVIÇO DO BRASIL ★ ★ ★ WWW.FOLHA.COM.BR

FOLHA DE S.PAULO

DIRETOR DE REDAÇÃO: OTAVIO FRIAS FILHO

DOMINGO, 5 DE ABRIL DE 2009
ANO 89 ★ Nº 29.222

EDIÇÃO SÃO PAULO, CONCLUÍDA ÀS 22H01 ★ R$ 4,00

ARQUIVO PÚBLICO DO ESTADO DE SÃO PAULO • BIBLIOTECA E HEMEROTECA

história

Ficha de Dilma Rousseff no Dops

CAPTURADO

Grupo de **Dilma** planejou sequestro de **Delfim Netto**

FERNANDA ODILA
DA SUCURSAL DE BRASÍLIA

Antonio Spinoza, ex-colega da ministra (Casa Civil), diz que o grupo armado que dirigiram teve como alvo o titular da Fazenda em 1969.

A ação chegou a ter data e local definidos. Um mapa da emboscada consta de processo no STM. Dilma , hoje aliada de Delfim, negou de forma "peremptória". "Ele fantasiou. Não me lembro disso." Pág. A8

In 2009, the current Brazilian president was just a pre-candidate (for the 2010 presidential election).

How it started

In 2009, the current Brazilian president was just a pre-candidate (for the 2010 presidential election).

The newspaper *A Folha de SP* (think New York Times) ran article about her.

How it started


UM JORNAL A SERVIÇO DO BRASIL ★ ★ ★ WWW.FOLHA.COM.BR

FOLHA DE S.PAULO

DIRETOR DE REDAÇÃO: OTAVIO FRIAS FILHO

DOMINGO, 5 DE ABRIL DE 2009
ANO 89 ★ Nº 29.222

EDIÇÃO SÃO PAULO, CONCLUÍDA ÀS 22H01 ★ R$ 4,00

ARQUIVO PÚBLICO DO ESTADO DE SÃO PAULO • BIBLIOTECA HERALDO

História

Ficha de Dilma Rousseff no Dops

CAPTURADO

Grupo de Dilma planejou sequestro de Delfim Netto

FERNANDA ODILA
DA SUCURSAL DE BRASÍLIA

Antonio Spinoza, ex-colega da ministra (Casa Civil), diz que o grupo armado que dirigiram teve como alvo o titular da Fazenda em 1969.

A ação chegou a ter data e local definidos. Um mapa da emboscada consta de processo no STM. Dilma, hoje aliada de Delfim, negou de forma "peremptória". "Ele fantasiou. Não me lembro disso." Pág. A8

In 2009, the current Brazilian president was just a pre-candidate (for the 2010 presidential election).

The newspaper *A Folha de SP* (think New York Times) ran article about her.

It had a "scan of her criminal files" as political activist in the military dictatorship period (1964-1985).

MEMÓRIA DA DITADURA

DILMA ROUSSEFF MINISTRA DA CASA CIVIL

Aos 19, 20 anos, achava que eu estava salvando o mundo

Dilma diz não ter a mesma cabeça da época em que era guerrilheira, mas se orgulha de não ter mudado de lado, e sim de métodos

FERNANDA ODILLA
DA SUCURSAL DE BRASÍLIA

UMA DAS três sentenças de prisão de Dilma Rousseff, de 1971, a descreve como a inimiga que "jamais esmoreceu" desde que ingressou na luta armada contra o regime instalado pelo golpe de 31 de março de 1964 e dissolvido 21 anos depois. Leia a entrevista da ministra sobre a vida na clandestinidade durante a ditadura.

FOLHA - A sra. se lembra dos planos para sequestrar Delfim e montar fábrica de explosivos?

DILMA ROUSSEFF - Ah, pelo amor de Deus. Nenhuma das duas eu lembro. Nunca ninguém do Exército, da Marinha e da Aeronáutica me perguntou isso. Não sabia disso. Acho que não era o que a gente [queria], não era essa a posição da VAR.

FOLHA - A sra. logo percebeu que a clandestinidade seria o caminho natural?

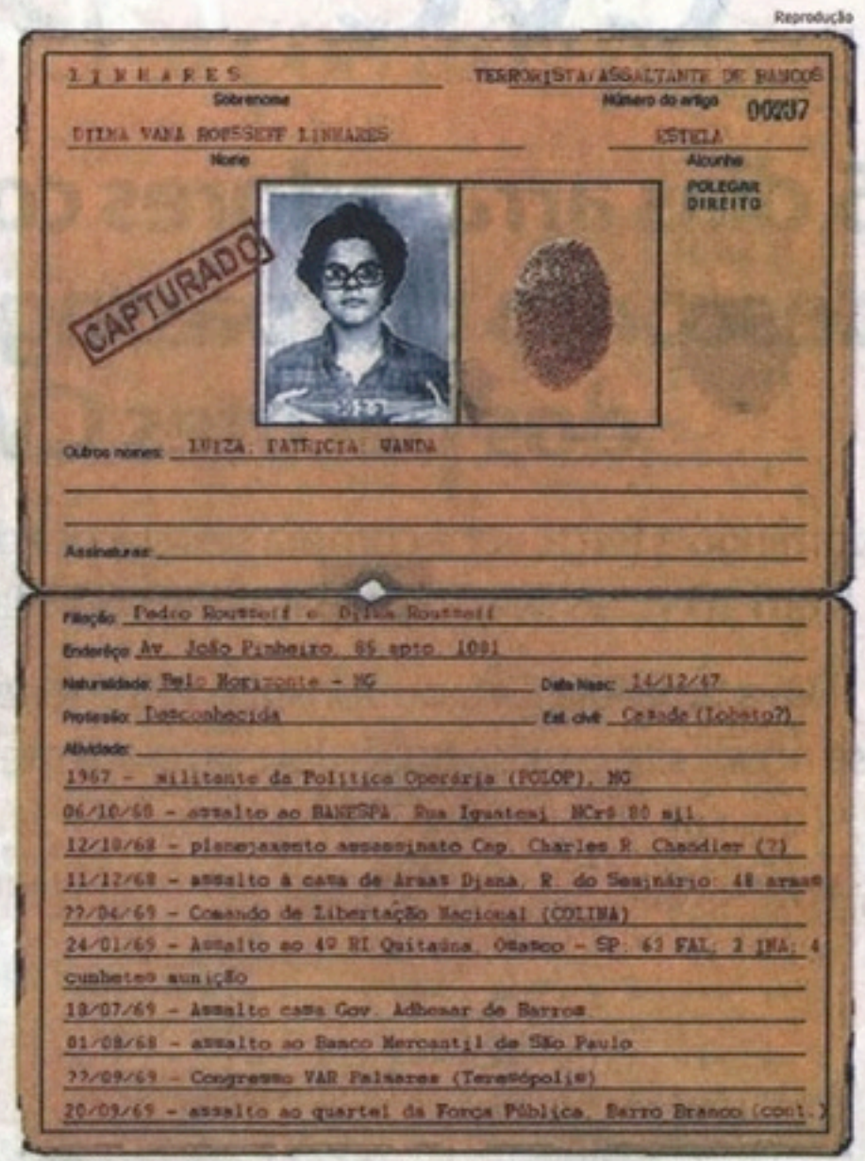
DILMA - Percebi. Todo mundo achava que podia haver no Brasil algo muito terrível. O receio de que um dia eles amanheceriam e começariam a matar era muito forte. Sou bem velha, comecei em 1964. Com o passar do tempo, o Brasil foi se fechando, as coisas foram ficando cada vez mais qualificadas como subversivas. Era subversivo até uma música, uma peça de teatro, qualquer manifestação de rua. Discutir reforma universitária era subversivíssimo. Coisas absolutamente triviais hoje eram muito subversivas.

dos nós. Não mudei de lado não, isso é um orgulho. Mudei de métodos, de visão. Inclusive, por causa daquilo, eu entendi muito mais coisas.

FOLHA - Como o quê?

DILMA - O valor da democracia, por exemplo. Por causa daquilo, eu entendi os processos absolutamente perversos. A tortura é um ato perverso. Tem um componente da tortura que é o que fizeram com aqueles meninos, os arrependidos, que iam para a televisão. Além da tortura, você tira a honra da pessoa. Acho que fizeram muito isso no Brasil. Por isso, minha filha, esse seu jornal não pode chamar a ditadura de ditabranda, viu? Não pode, não. Você não sabe o que é a quantidade de secreção que sai de um ser humano quando ele apanha e é torturado. Porque essa quantidade de líquidos que nós temos, o sangue, a urina e as fezes aparecem na sua forma mais humana. Não dá para chamar isso de ditabranda, não.

Reprodução



Ficha de Dilma após ser presa com crimes atribuídos a ela, mas que ela não cometeu

Oban e um mês no Dops. Eu custei a ir embora da Oban. Achava estranho eu não ir embora. Todo mundo ia, e eu ficava. Eu não lembro a data. Vai ficando muito obscuro, como foi e como é que não foi.

FOLHA - Vocês passavam por um treinamento intensivo para deletar as coisas. Tinha que esquecer para não contar?

DILMA - Uma parte você tentava esquecer. Sabe que teve uma época em que eu falei uma coisa que eu achava que era verdade e não era. Era mentira que eu tinha contado e aí depois eu descobri que era mentira. Você conta e se convence.

FOLHA - Informação obtida sob tortura é de responsabilidade de quem tortura e não de quem fala? Dá para culpar a pessoa que falou?

DILMA - Não dá mesmo. Até porque ali, naquela hora, tinha uma coisa muito engraçada que eu vi. Aconteceu com muita gente, não foi só comigo. É por isso que aquela pergunta é absurda, a do senador [Agripino Maia, do DEM]. A mentira é uma imensa vitória e a verdade é a derrota. Na chegada do presídio [Tiradentes], estava escrito "Feliz do povo que não tem heróis", que era uma frase do Brecht que tem um sentido amplo. Esse fato de não precisar de heróis mostra uma grande civilidade. É preciso que cada um tenha um pouco de heroísmo.

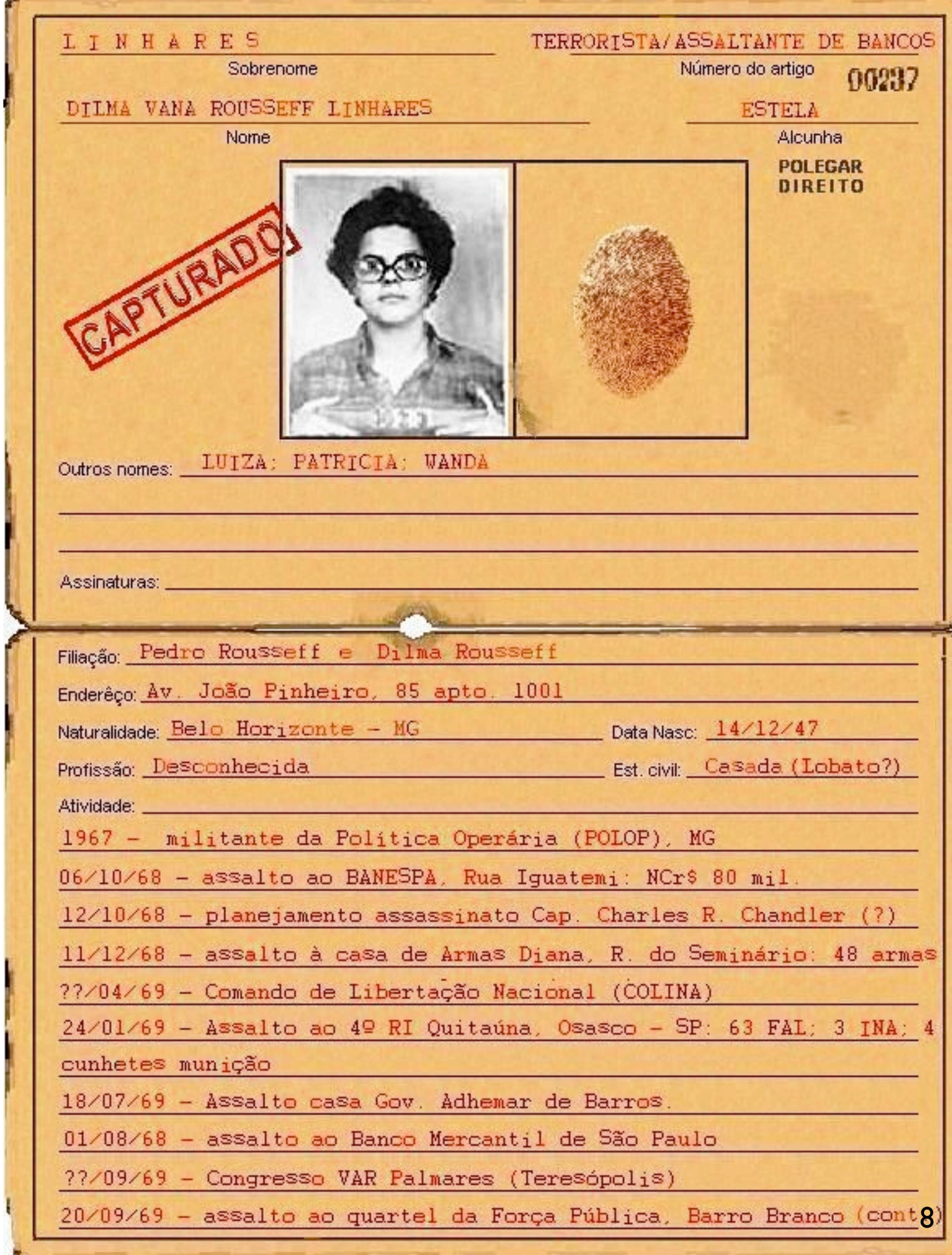
FOLHA - Quando a sra. chegou à Oban, houve muitos gritos?

DILMA - Teve. Fazia parte do script. É uma luta eterna entre a sua autodestruição e sua luta

Criminal Records?

A "scan" of her personal files maintained by the military internal security during the Brazilian military regime.

The Public Archive of SP actually hosts such a collection.


L I N H A R E S — Sobrenome

TERRORISTA/ASSALTANTE DE BANCOS — Número do artigo 00237

DILMA VANA ROUSSEFF LINHARES — Nome

ESTELA — Alcunha

POLEGAR DIREITO

CAPTURADO

Outros nomes: LUIZA; PATRICIA; WANDA

Assinaturas:

Filiação: Pedro Rousseff e Dilma Rousseff

Endereço: Av. João Pinheiro, 85 apto. 1001

Naturalidade: Belo Horizonte - MG — Data Nasc: 14/12/47

Profissão: Desconhecida — Est. civil: Casada (Lobato?)

Atividade:

1967 - militante da Política Operária (POLOP), MG

06/10/68 - assalto ao BANESPA, Rua Iguatemi: NCr$ 80 mil.

12/10/68 - planejamento assassinato Cap. Charles R. Chandler (?)

11/12/68 - assalto à casa de Armas Diana, R. do Seminário: 48 armas

??/04/69 - Comando de Libertação Nacional (COLINA)

24/01/69 - Assalto ao 4º RI Quitaúna, Osasco - SP: 63 FAL; 3 INA; 4 cunhetes munição

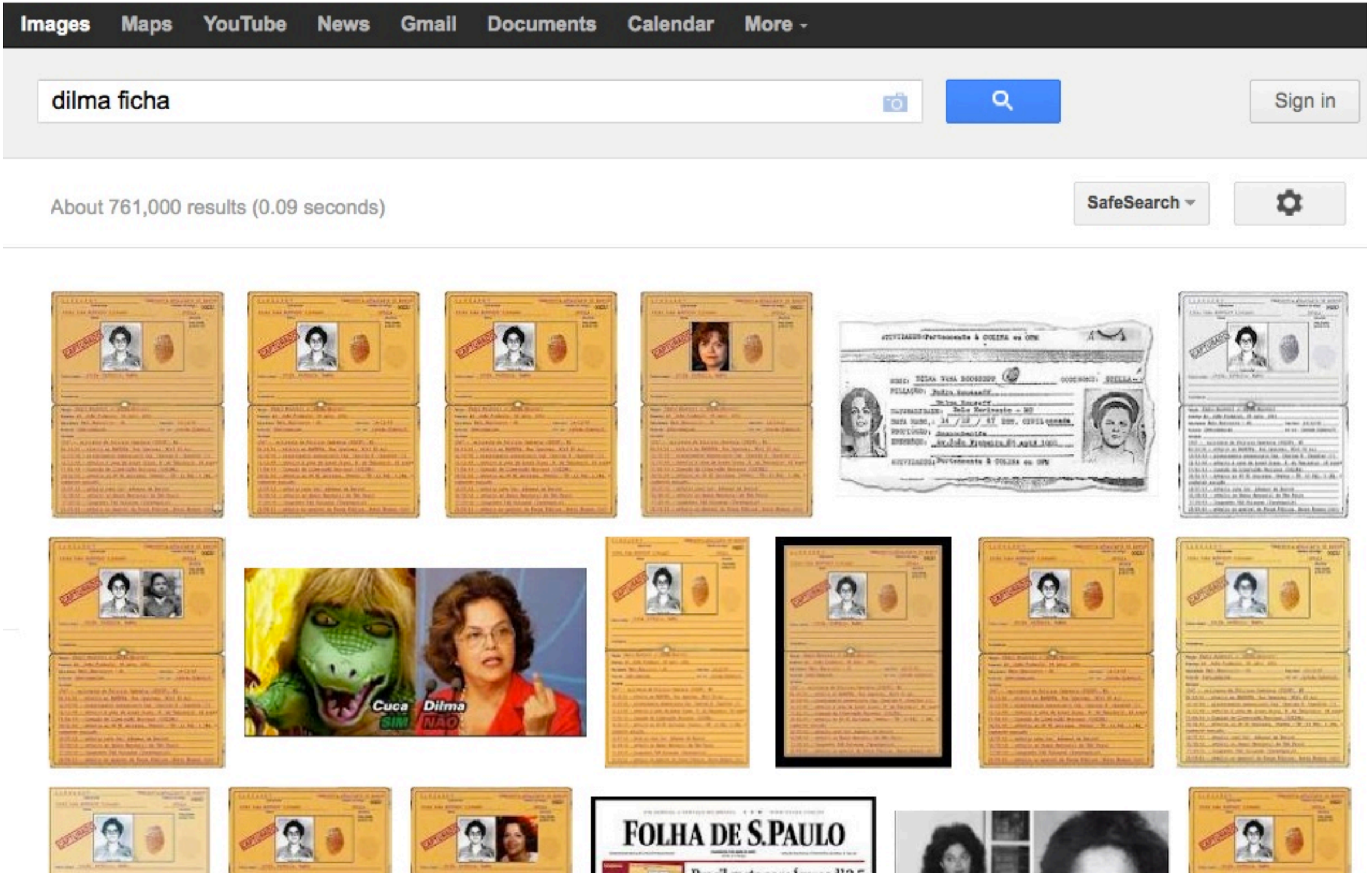
18/07/69 - Assalto casa Gov. Adhemar de Barros.

01/08/68 - assalto ao Banco Mercantil de São Paulo

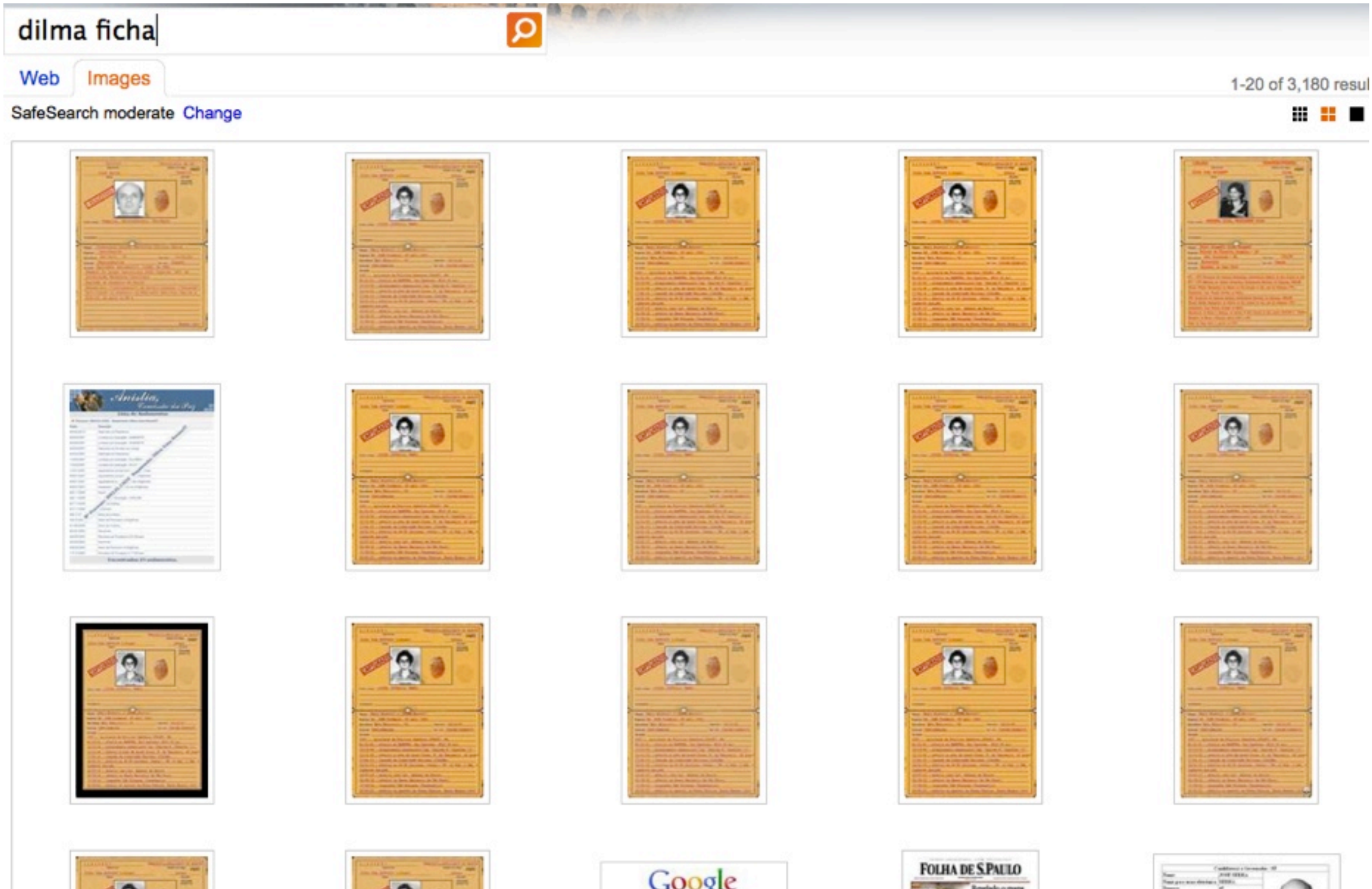
??/09/69 - Congresso VAR Palmares (Teresópolis)

20/09/69 - assalto ao quartel da Força Pública, Barro Branco (cont

Searching the Web...



Searching the Web...



Criminal Records?

Criminal Records?

- This image was already going around the net for ≈ six months − it is a clear fake.

Criminal Records?

- This image was already going around the net for ≈ six months − it is a clear fake.
- She hired us, as consultants, to provide a forensic analysis of its authenticity that could hold on court.

Criminal Records?

- This image was already going around the net for ≈ six months − it is a clear fake.
- She hired us, as consultants, to provide a forensic analysis of its authenticity that could hold on court.
- There were several versions of the image (near duplicates)
  → Which one was the original?
  → Where should we perform the analysis?

How to find the Original?

- The images are “copied” around...
  - resized;
  - cropped;
  - color corrected;
  - recompressed;
  - and possibly other transformations.

Media Phylogeny

- Identify, among a set of near duplications, which element is the original, and the structure of generation of each near duplication.

- Tells the history of the transformations created the duplications.

Media Phylogeny

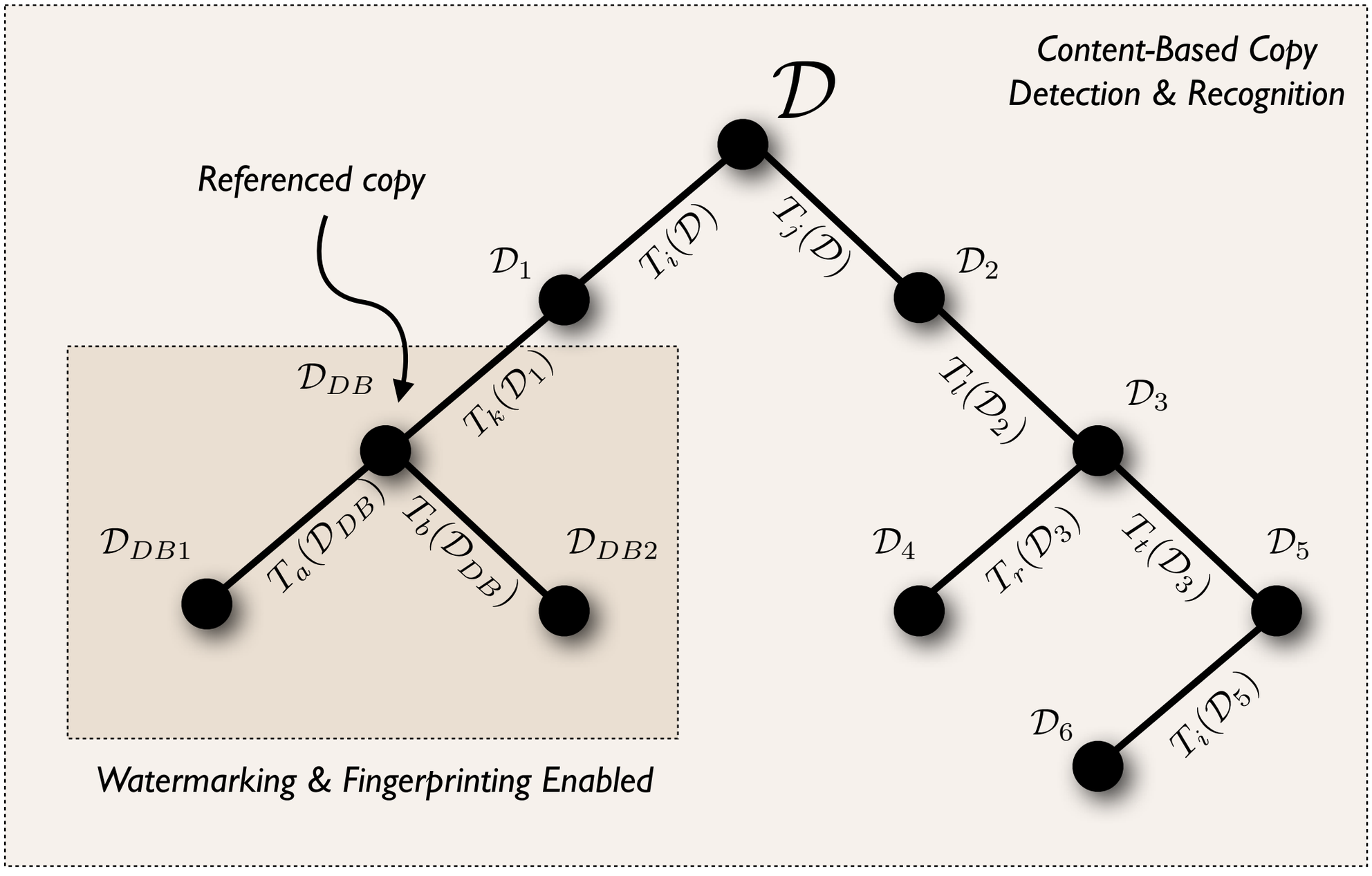


Image Phylogeny Trees

IPT

Image Phylogeny Trees: IPT

- **Security**.
- **Forensics**.
- **Copyright enforcement**.
- **News tracking services**.
- **Indexing**.

Image Phylogeny Trees: IPT

- **Security**: the modification graph provides information of suspects' behavior, and points out flow of content distribution.

- **Forensics**.
- **Copyright enforcement**.
- **News tracking services**.
- **Indexing**.

Image Phylogeny Trees: IPT

- **Security**.
- **Forensics**: analysis in the original document (root of the tree) instead of in a near duplicate.
- **Copyright enforcement**.
- **News tracking services**.
- **Indexing**.

Image Phylogeny Trees: IPT

- **Security**.
- **Forensics**.
- **Copyright enforcement**: traitor tracing without the need of source control techniques (watermarking or fingerprinting).
- **News tracking services**.
- **Indexing**.

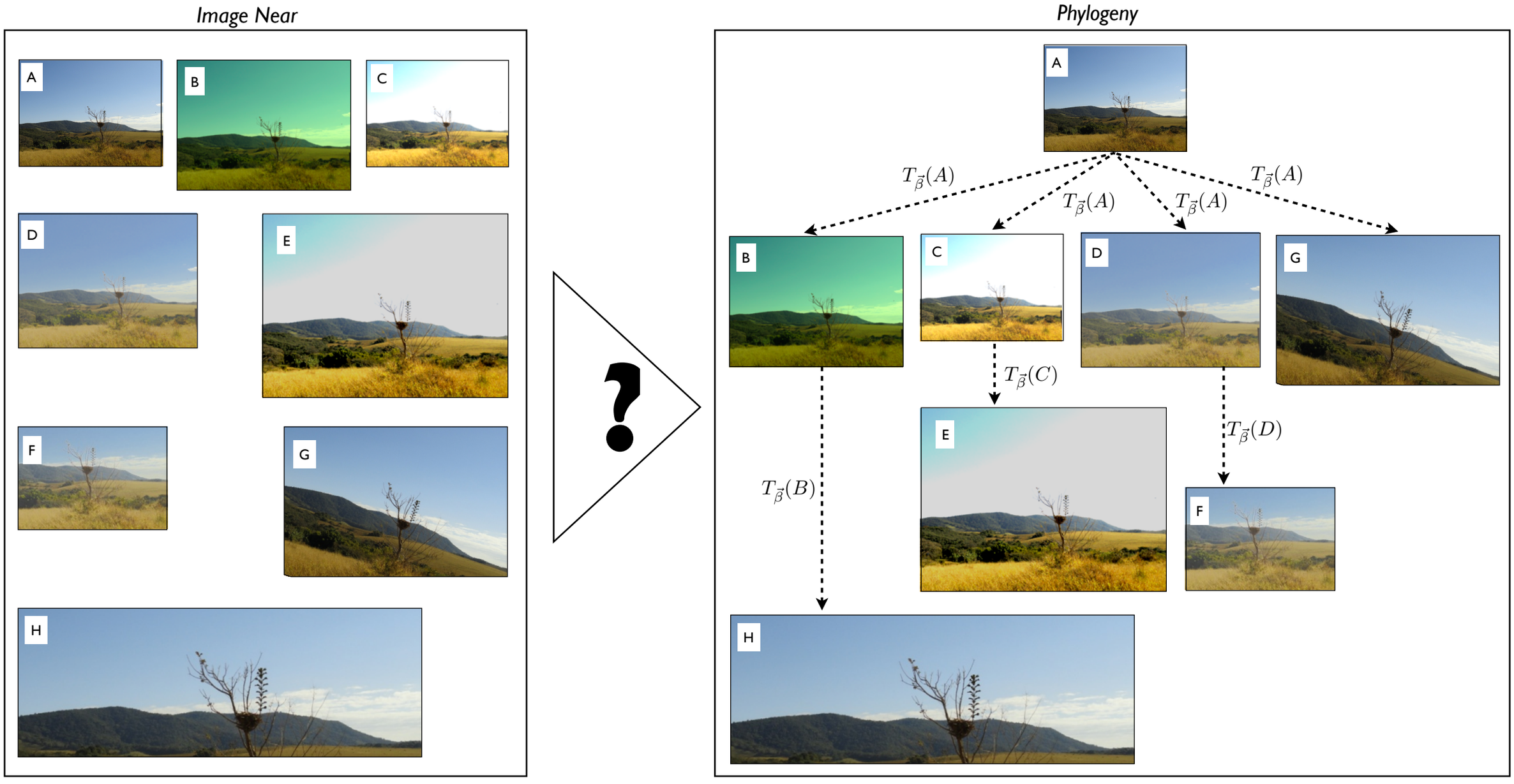
Image Phylogeny Trees: IPT

- **Security**.
- **Forensics**.
- **Copyright enforcement**.
- **News tracking services**: the ND relationships can feed news tracking services with key elements for determining the opinion forming process across time and space.
- **Indexing**.

Image Phylogeny Trees: IPT

- **Security**.
- **Forensics**.
- **Copyright enforcement**.
- **News tracking services**.
- **Indexing**: tree root can give us an image from an ND set as a representative to index, store, or even further refine the ND search.
  Tree structure might help indexing and retrieving.

Our Objective



Two Subproblems

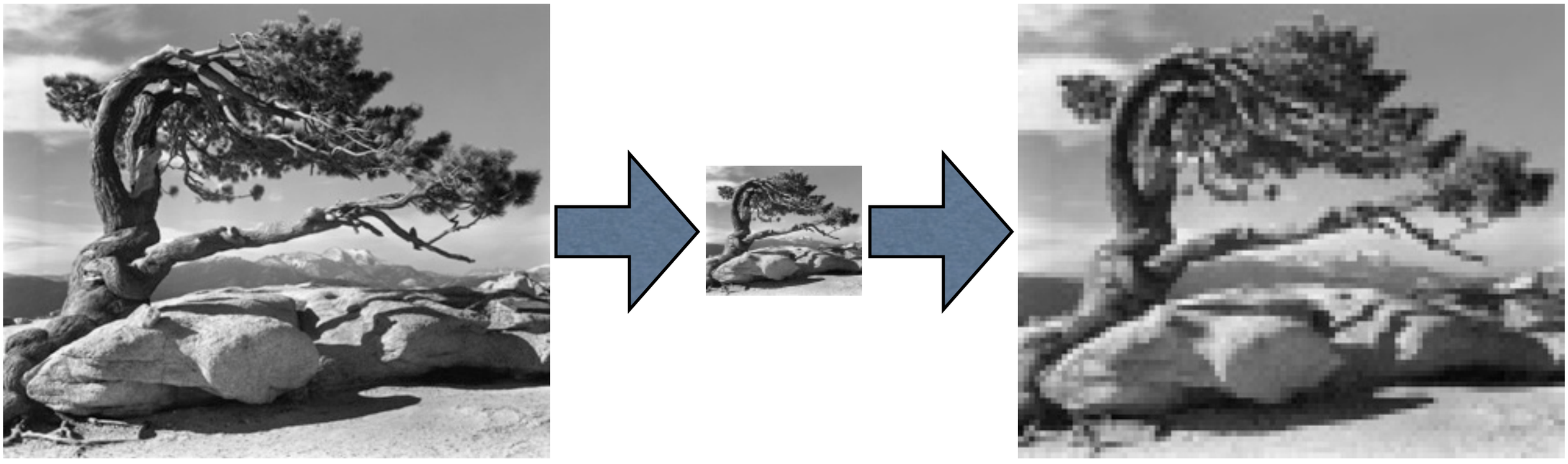
1. Define good dissimilarity functions $d(i, j)$ between images.

2. Develop algorithms that construct the Image Phylogeny Tree given a dissimilarity matrix of the images.

Dissimilarity

Dissimilarity

The dissimilarity is not a metric - we want to estimate how likely A→B and B→A.



Think about cropping, or resizing an image - these are not two-way operations.

Jeffrey Pine, Sentinel Dome, Yosemite National Park, **Ansel Adams**.

Dissimilarity

- Define a family of image transformations $T_\beta(I)$ parameterized by β.

- Let $d_\beta(i,j) = |I_j - T_\beta(I_i)|^2$
  find β_{min} that minimizes $d_\beta(i,j)$

$$d(i,j) = d_{\beta_{min}}(i,j)$$

Dissimilarity

$$T_\beta(I) = T_{jpeg}(T_{color}(T_{spatial}(I))$$

- We use a composition of three simple steps.
- In the general case, finding the optimum parameters of a general transformation might be a complicated optimization.

Dissimilarity

Dissimilarity

- Spatial
  - Affine Transformation
  - Cropping

Dissimilarity

- Spatial
  - Affine Transformation
  - Cropping
- Color
  - Channel Brightness and Contrast.

Dissimilarity

- Spatial
  - Affine Transformation
  - Cropping
- Color
  - Channel Brightness and Contrast.
- JPEG Compression
  - Quantization tables.

Spatial Transformation for the Dissimilarity

Afghanistan-Pakistan border, **Steve McCurry**.

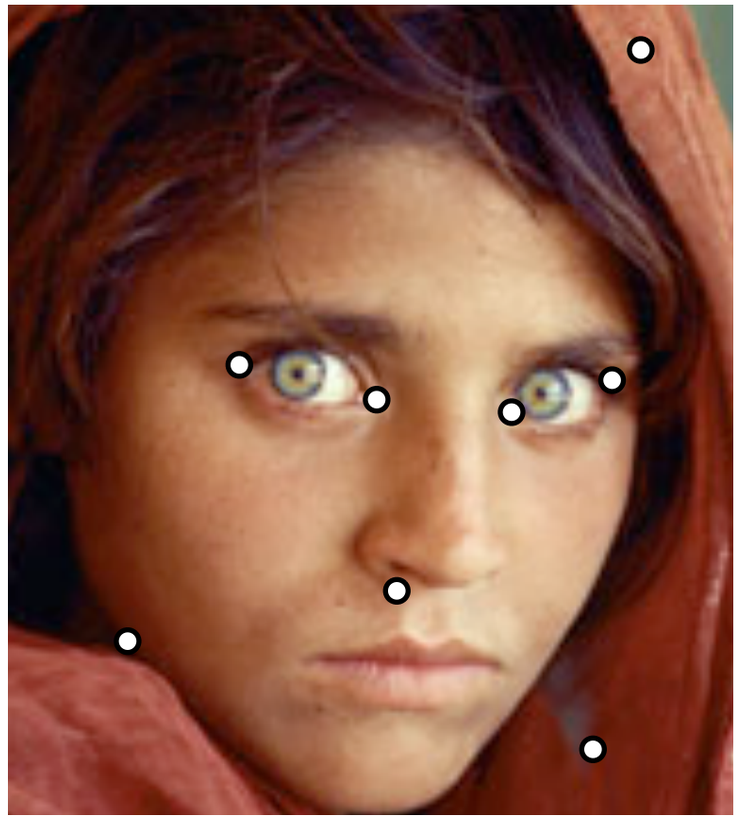
Spatial Transformation for the Dissimilarity

Key Points.

Afghanistan-Pakistan border, **Steve McCurry**.

Spatial Transformation for the Dissimilarity

Key Points.

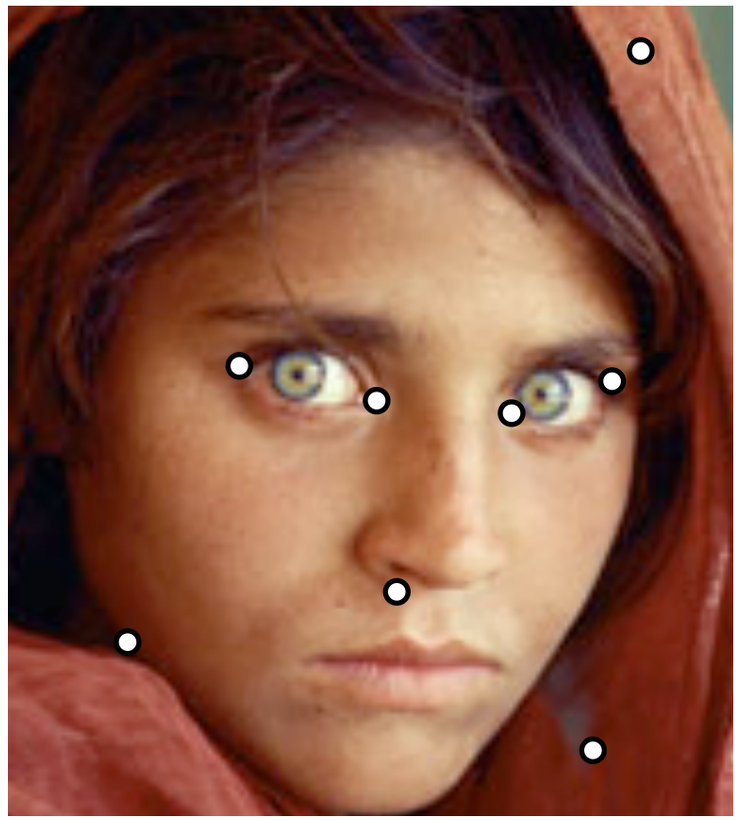


Afghanistan-Pakistan border, **Steve McCurry**.

Spatial Transformation for the Dissimilarity

Key Points.

Correspondences.



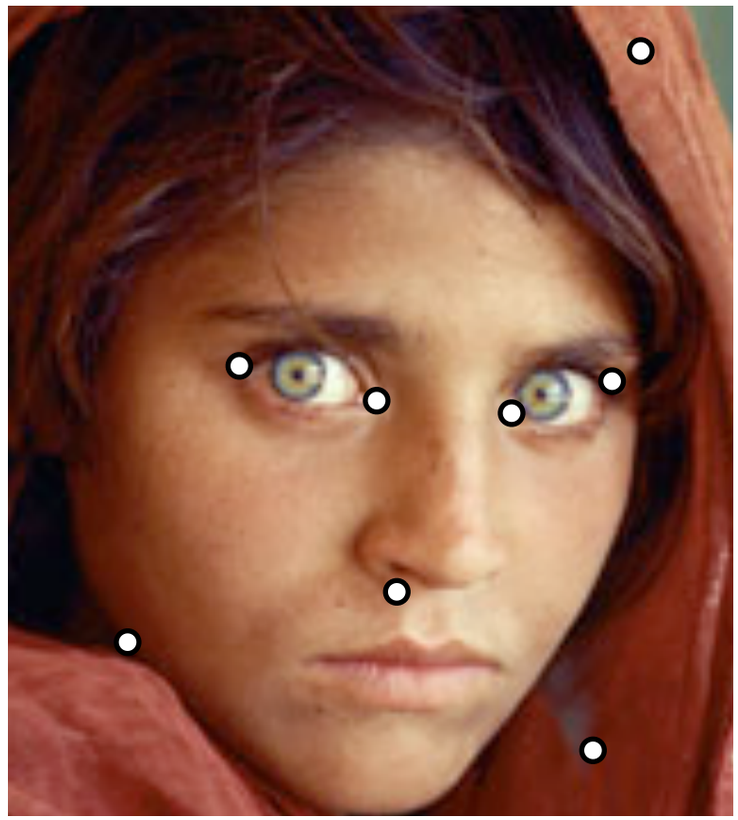
Afghanistan-Pakistan border, **Steve McCurry**.

Spatial Transformation for the Dissimilarity

Key Points.

Correspondences.

Robust Estimation
of Affine Transf.



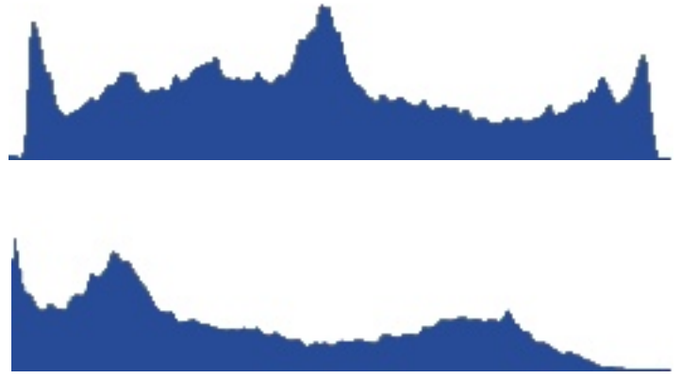
$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

Afghanistan-Pakistan border, **Steve McCurry**.

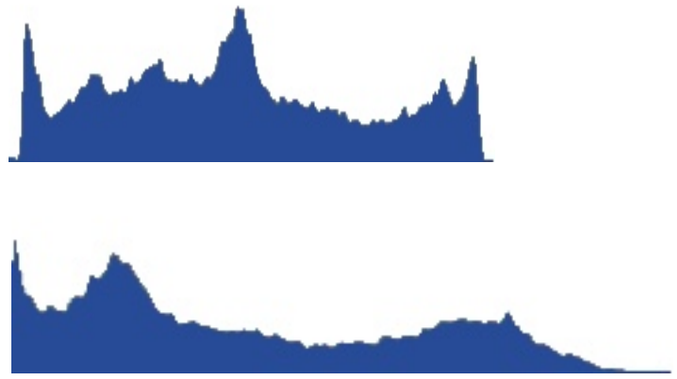
Color Transformation for the Dissimilarity

V-J Day, Times Square, **Alfred Eisenstaedt**

Color Transformation for the Dissimilarity



Color Transformation for the Dissimilarity



Brightness / Contrast ~ Same mean and stddev.

Dissimilarity: Compression

- Use the quantization table of the jpeg of B to compress T(A).

Tree Reconstruction

Tree Construction

Tree Construction

- Local decisions of direction on pairs of images is not a good idea...

Tree Construction

- Local decisions of direction on pairs of images is not a good idea...
- Proposition: we want a MST.

Tree Construction

- Local decisions of direction on pairs of images is not a good idea...
- Proposition: we want a MST.

  ...but we have a complete directed graph.

MST of directed graphs in the Literature

The *Optimum Branching* problem finds the MST of a directed graph for a given root.

In our context, it would have to be applied to each vertex as a root, and the final complexity in our scenario would be O(n^3).

It also uses a Fibonacci Heap.

Oriented Kruskal

Algorithm 1 Oriented Kruskal

Require: a dissimilarity matrix M

1: **for** $i \in [1..n]$ **do** ▷ Initialization
2: $Parent[i] \leftarrow i$
3: **end for**
4: $Sorted \leftarrow$ sort positions (i, j) of M into nondecreasing order
5: $n_{edges} \leftarrow 0$ ▷ Controls stopping criterium
6: **for** each position $(i, j) \in Sorted$ **do**
7: **if** $(\texttt{Root}(i) \neq \texttt{Root}(j))$ **then** ▷ **Test I:** joins different trees
8: **if** $(Root(j) = j)$ **then** ▷ **Test II:** endpoint must be a root
9: $Parent[j] \leftarrow i$
10: $n_{edges} \leftarrow n_{edges} + 1$
11: **end if**
12: **end if**
13: **if** $(n_{edges} = n - 1)$ **then** ▷ The IPT has already n-1 edges
14: **return** $Parent$ ▷ Returning the final IPT
15: **end if**
16: **end for**

Oriented Kruskal

Our method runs once, and finds both the root and structure simultaneously.

It has an $O(n^2 \log n)$ complexity – we need to sort all n^2 edges of the complete graph.

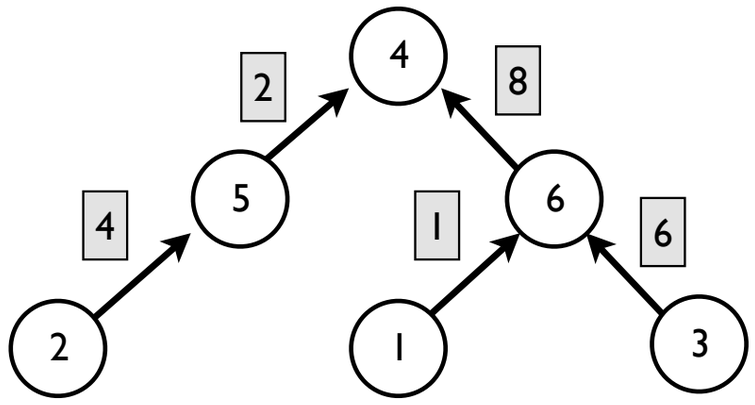
It requires the Union-Find data structure.

Construction Example

Dissimilarity Matrix

M	1	2	3	4	5	6
1	-	31	57	37	45	49
2	31	-	33	23	29	32
3	51	41	-	42	37	38
4	16	36	28	-	15	27
5	35	18	54	30	-	54
6	12	40	22	60	19	-

Reconstructed Tree [6, 5, 6, 4, 4, 4]



Algorithm Steps

✔	1	M[6,1] =	12	Select Edge (1 → 6)
✔	2	M[4,5] =	15	Select Edge (5 → 4)
×	3	M[4,1] =	16	**Test II:** Root(1) = 6
✔	4	M[5,2] =	18	Select Edge (2 → 5)
×	5	M[6,5] =	19	**Test II:** Root(5) = 4
✔	6	M[6,3] =	22	Select Edge (3 → 6)
×	7	M[2,4] =	23	**Test I:** Root(2) = Root(4)
✔	8	M[4,6] =	27	Select Edge (6 → 4)

Evaluation: Comparing Trees

$$\textbf{Root: } R(\mathrm{IPT}_1, \mathrm{IPT}_2) = \begin{cases} 1, & \text{If } \texttt{Root}(\mathrm{IPT}_1) = \texttt{Root}(\mathrm{IPT}_2) \\ 0, & \text{Otherwise} \end{cases}$$

Evaluation: Comparing Trees

$$\textbf{Root: } R(\mathrm{IPT}_1, \mathrm{IPT}_2) = \begin{cases} 1, & \text{If } \texttt{Root}(\mathrm{IPT}_1) = \texttt{Root}(\mathrm{IPT}_2) \\ 0, & \text{Otherwise} \end{cases}$$

$$\textbf{Edges: } E(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|E_1 \cap E_2|}{n-1}$$

Evaluation: Comparing Trees

Root: $R(\mathrm{IPT}_1, \mathrm{IPT}_2) = \begin{cases} 1, & \text{If } \texttt{Root}(\mathrm{IPT}_1) = \texttt{Root}(\mathrm{IPT}_2) \\ 0, & \text{Otherwise} \end{cases}$

Edges: $E(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|E_1 \cap E_2|}{n-1}$

Leaves: $L(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$

Evaluation: Comparing Trees

Root: $R(\mathrm{IPT}_1, \mathrm{IPT}_2) = \begin{cases} 1, & \text{If } \texttt{Root}(\mathrm{IPT}_1) = \texttt{Root}(\mathrm{IPT}_2) \\ 0, & \text{Otherwise} \end{cases}$

Edges: $E(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|E_1 \cap E_2|}{n-1}$

Leaves: $L(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$

Ancestry: $A(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$

IPT Experiments

IPT Experiments

- Experimental Setup.

IPT Experiments

- Experimental Setup.
- Complete Trees.

IPT Experiments

- Experimental Setup.
- Complete Trees.
- Missing Nodes.

IPT Experiments

- Experimental Setup.
- Complete Trees.
- Missing Nodes.
  - Missing Root.

IPT Experiments

- Experimental Setup.
- Complete Trees.
- Missing Nodes.
  - Missing Root.
  - Missing Internal Nodes.

IPT Experiments

- Experimental Setup.
- Complete Trees.
- Missing Nodes.
  - Missing Root.
  - Missing Internal Nodes.
- Real ND sets from the Web.

IPT Experiments

- Experimental Setup.
- Complete Trees.
- Missing Nodes.
  - Missing Root.
  - Missing Internal Nodes.
- Real ND sets from the Web.
- A first look into Forests.

Experimental Setup

Experimental Setup

- 50 raw images from UCID.

Experimental Setup

- 50 raw images from UCID.
- Trees with 10, 20, 30, 40, and 50 nodes.

Experimental Setup

- 50 raw images from UCID.
- Trees with 10, 20, 30, 40, and 50 nodes.
- For every size, 50 random tree topologies, each with 10 different random parameters.

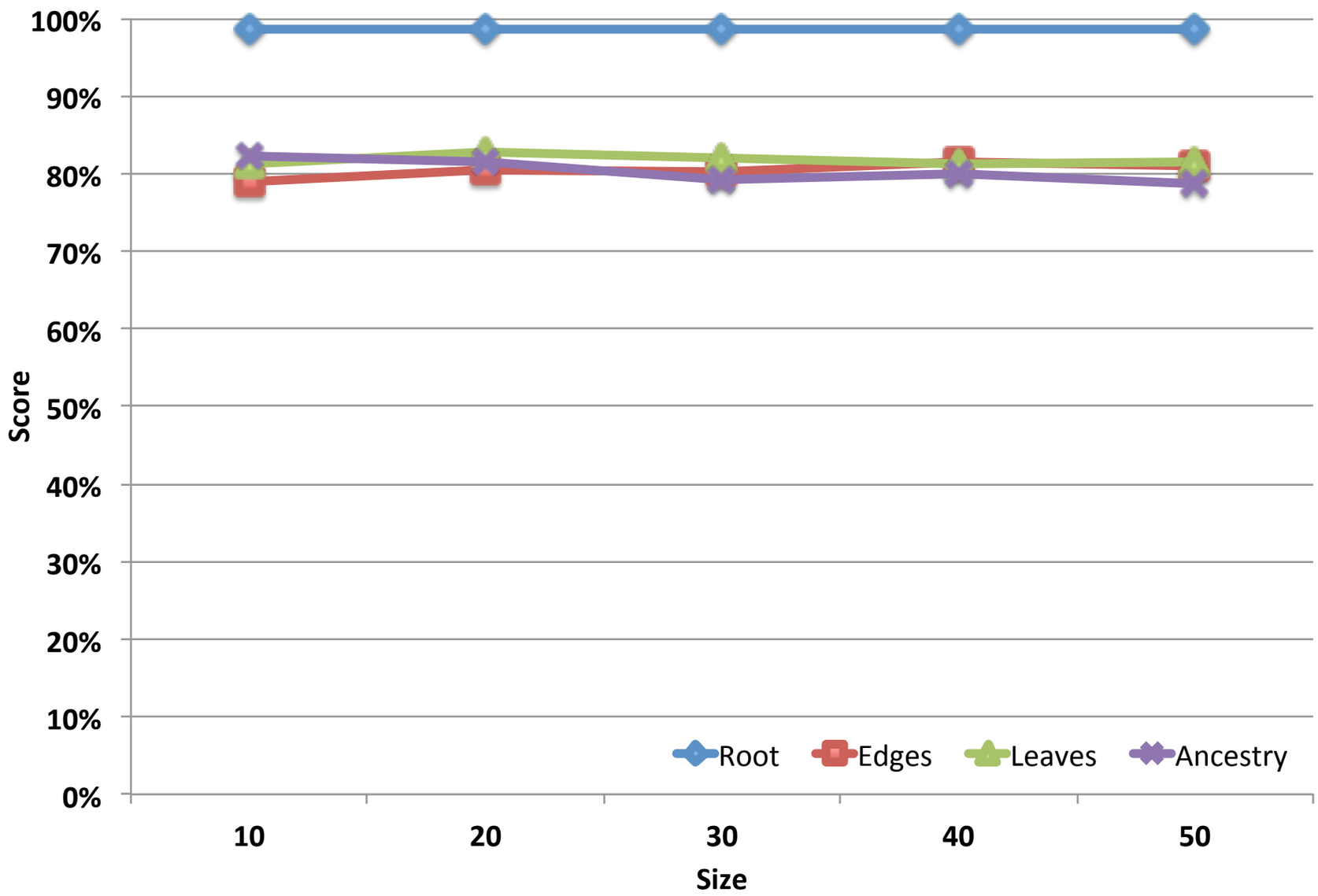
Experimental Setup

- 50 raw images from UCID.
- Trees with 10, 20, 30, 40, and 50 nodes.
- For every size, 50 random tree topologies, each with 10 different random parameters.
- ND set created with affine transformation, crop, brightness-contrast-gamma on each channel and compression. We use ImageMagick.

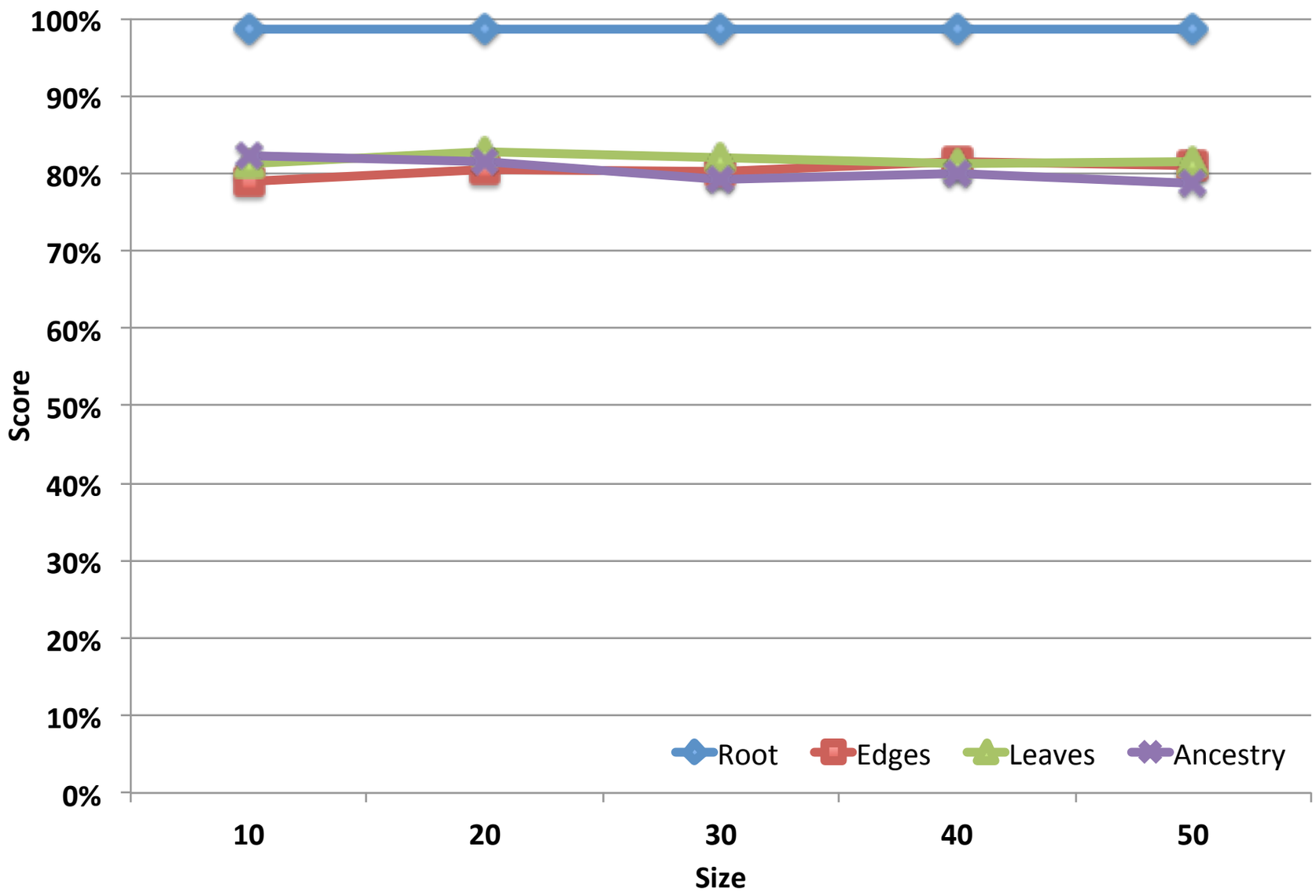
Experimental Setup

- 50 raw images from UCID.
- Trees with 10, 20, 30, 40, and 50 nodes.
- For every size, 50 random tree topologies, each with 10 different random parameters.
- ND set created with affine transformation, crop, brightness-contrast-gamma on each channel and compression. We use ImageMagick.
- Dissimilarity construction with OpenCV and libjpg: affine transformation, brightness-contrast by channel, and compression.

Complete Trees



Complete Trees

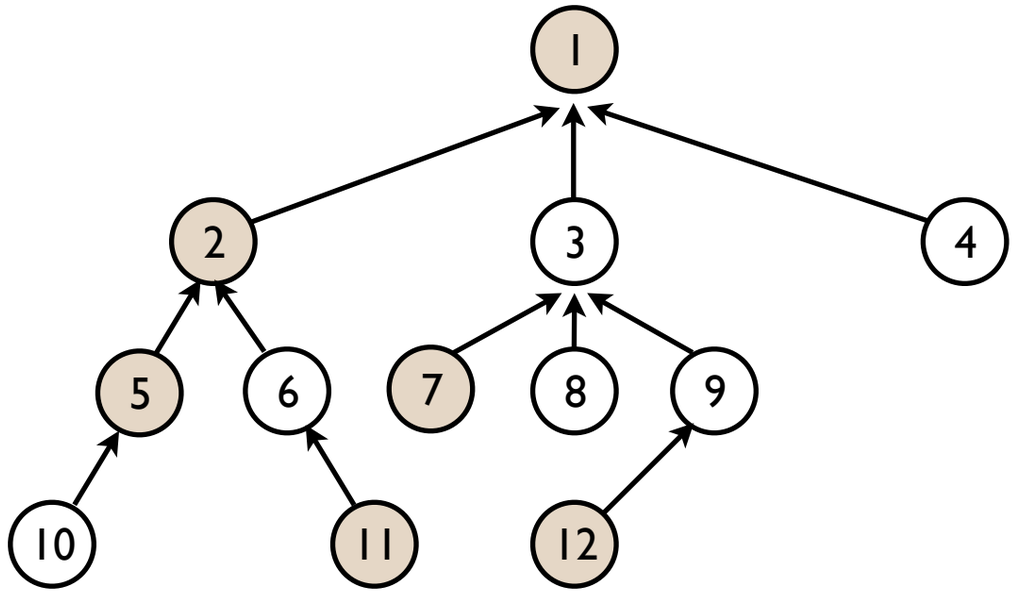


If the correct root is at depth zero, we identified the root of the tree. Here, regardless of the tree size, the average depth at which our solution finds the correct root is lower than 0.03.

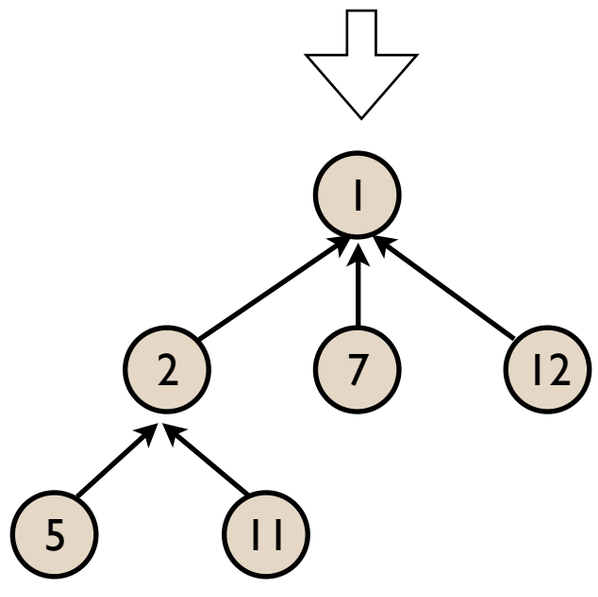
Missing Links

- On the wild, it is unrealistic to expect to have all the nodes of the tree.
- How to handle missing links?
  How do we evaluate the algorithm?

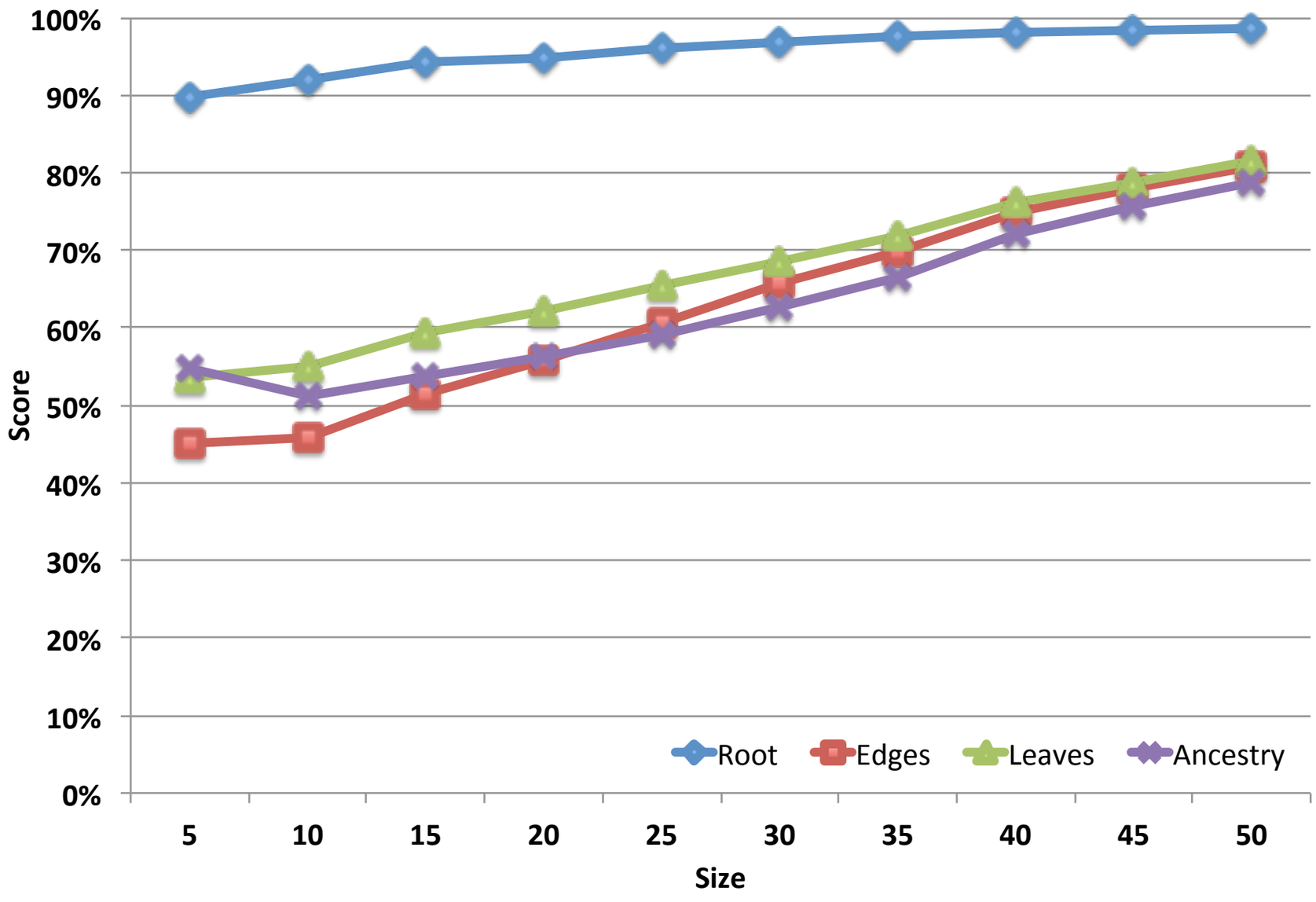
Missing Nodes



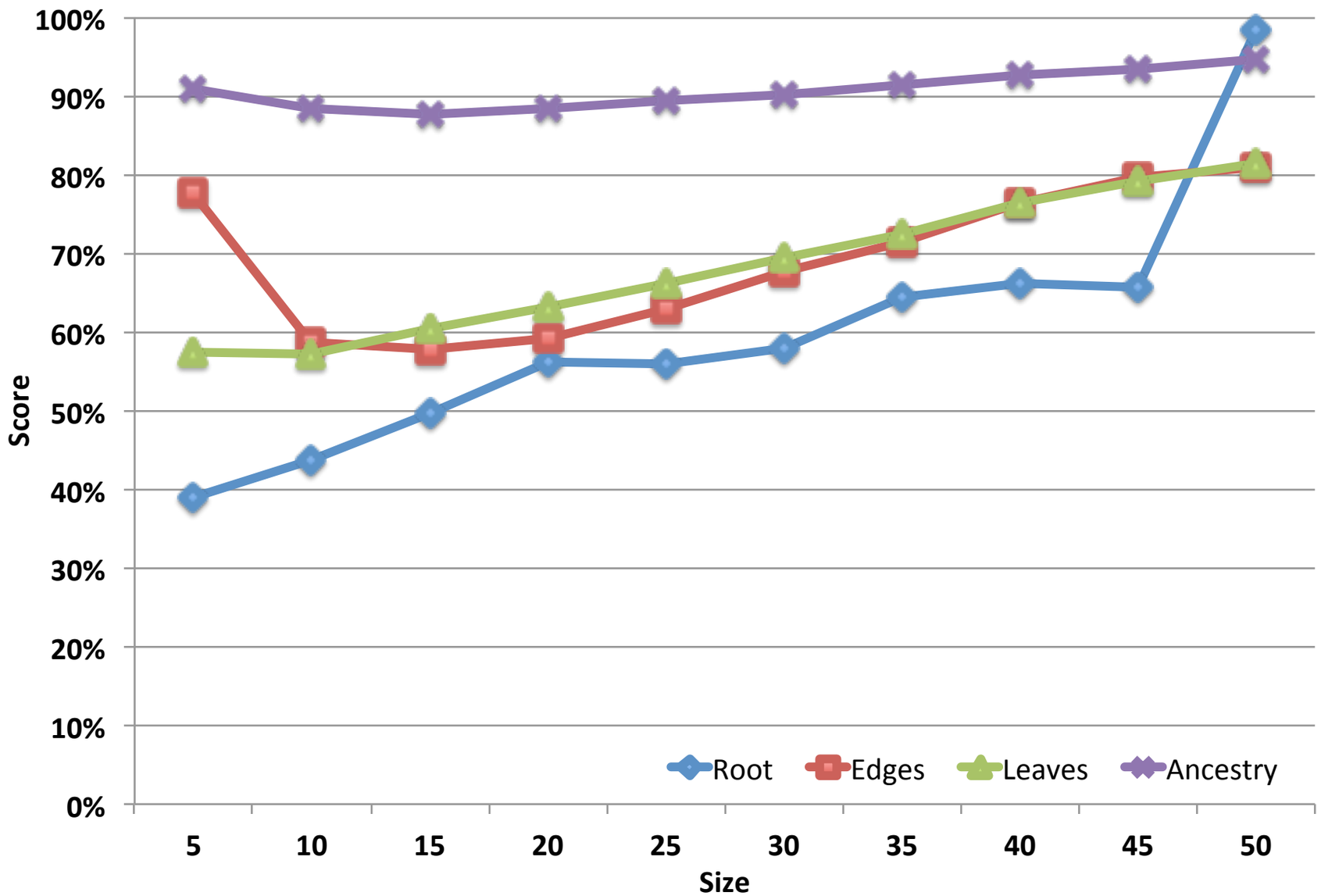
Using Ancestry Information



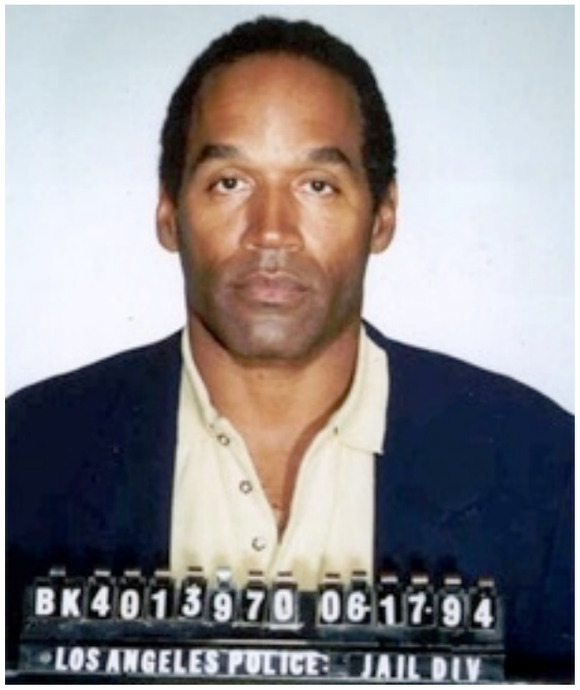
Missing only Internal Nodes

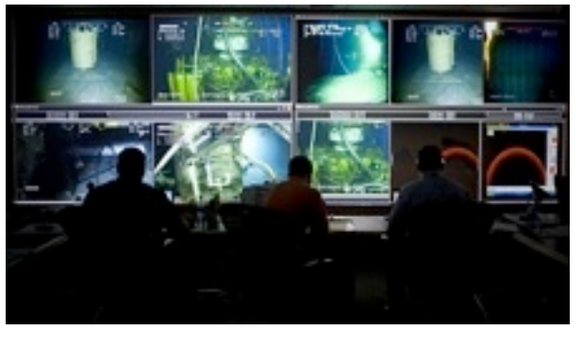


Missing Root and Internal Nodes



Real ND sets from Web





How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

Er1: one if the new node IB is not a child of its generating node IA.

How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

Er1: one if the new node IB is not a child of its generating node IA.

Er2: one if the structure of the tree relating the nodes in the original set changes with the insertion of the new node IB in the set.

How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

Er1: one if the new node IB is not a child of its generating node IA.

Er2: one if the structure of the tree relating the nodes in the original set changes with the insertion of the new node IB in the set.

Er3: one if the new node IB appears as a father of another node on the original tree.

How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

Er1: one if the new node IB is not a child of its generating node IA.

Er2: one if the structure of the tree relating the nodes in the original set changes with the insertion of the new node IB in the set.

Er3: one if the new node IB appears as a father of another node on the original tree.

Er4: is one if the root of the reconstructed tree of the original set is different from the root of the reconstructed tree of the set augmented with IB.

How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

Er1: one if the new node IB is not a child of its generating node IA.

Er2: one if the structure of the tree relating the nodes in the original set changes with the insertion of the new node IB in the set.

Er3: one if the new node IB appears as a father of another node on the original tree.

Er4: is one if the root of the reconstructed tree of the original set is different from the root of the reconstructed tree of the set augmented with IB.

P: one if the reconstructed tree is perfect compared to the original tree (Er1 = Er2 = 0).

How to Evaluate results?

- Since we do not know the ground truths, we evaluate the stability of reconstruction.

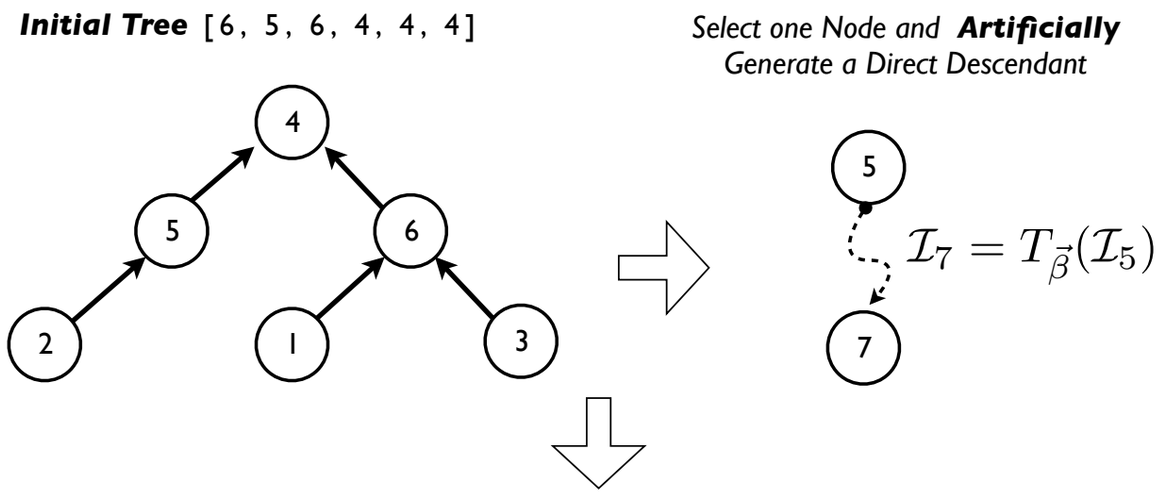
Er1: one if the new node IB is not a child of its generating node IA.

Er2: one if the structure of the tree relating the nodes in the original set changes with the insertion of the new node IB in the set.

Er3: one if the new node IB appears as a father of another node on the original tree.

Er4: is one if the root of the reconstructed tree of the original set is different from the root of the reconstructed tree of the set augmented with IB.

P: one if the reconstructed tree is perfect compared to the original tree (Er1 = Er2 = 0).



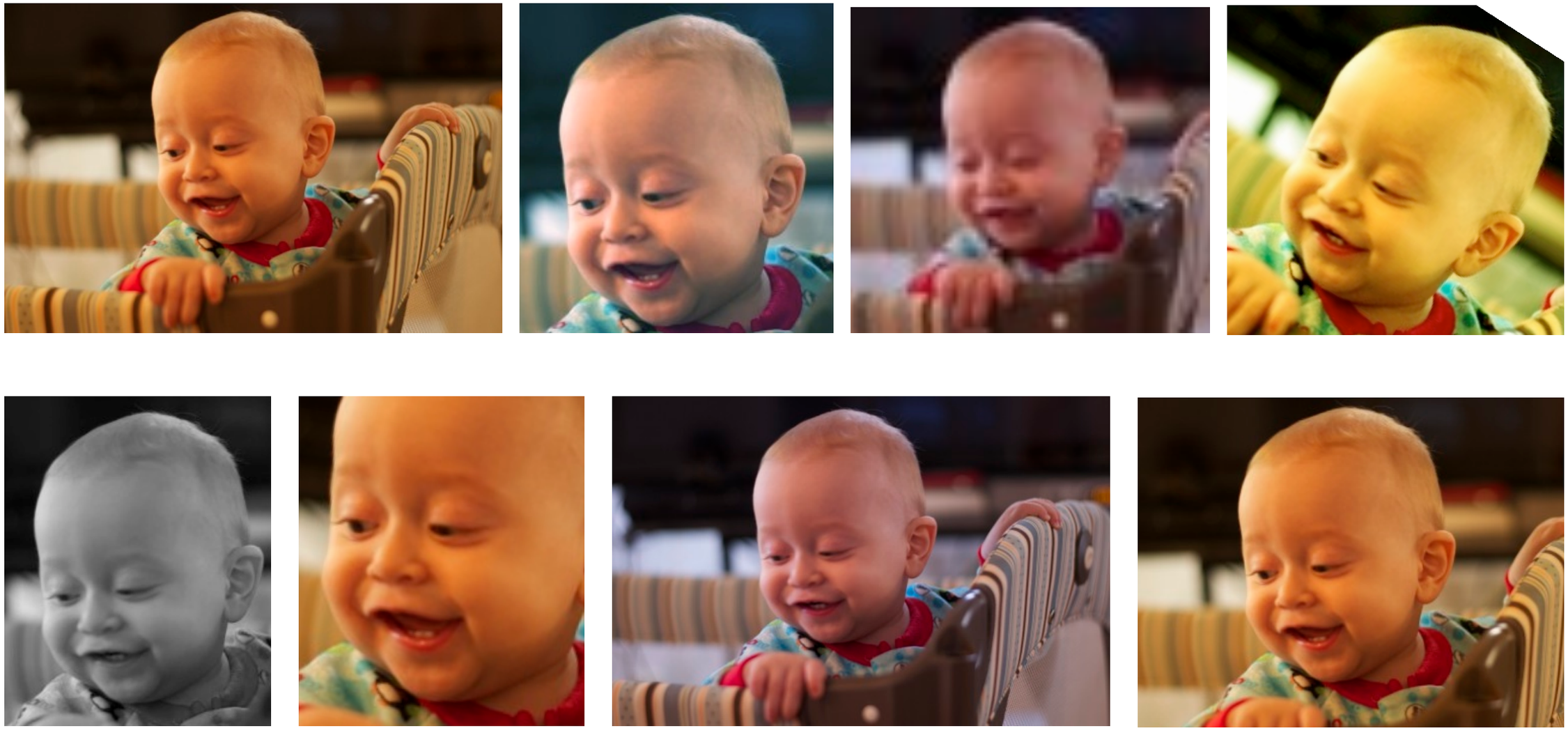
Reconstruction Errors	
Er_1	1
Er_2	1
Er_3	1
Er_4	0
Success	
P	0

Real ND sets

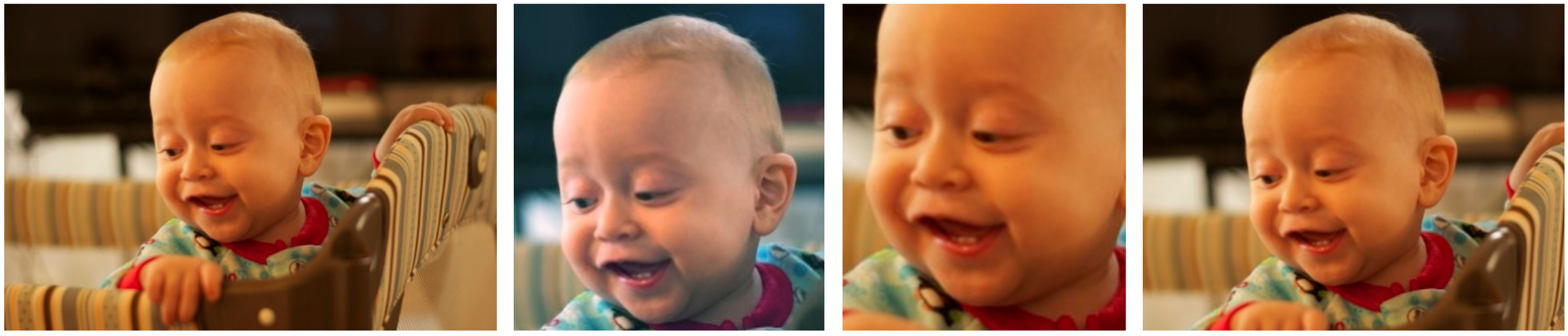
ORIENTED KRUSKAL IPT ALGORITHM RESULTS FOR THE UNCONSTRAINED SCENARIO.

	Description	# of Cases	$\%Er_1$	$\%Er_2$	$\%Er_3$	$\%Er_4$	%P
TG_1	Iranian Missiles	90	40.0%	11.1%	11.1%	0.0%	55.6%
TG_2	Bush Reading	95	17.9%	3.2%	3.2%	0.0%	81.1%
TG_3	WTC Tourist	95	25.3%	6.3%	6.3%	1.1%	71.6%
TG_4	BP Oil Spill	100	25.0%	0.0%	0.0%	0.0%	75.0%
TG_5	Israeli-Palestinian Peace Talks	95	21.1%	7.4%	7.4%	0.0%	75.8%
TG_6	Criminal Record	90	41.1%	13.3%	13.3%	0.0%	54.4%
TG_7	Palin and Rifle	100	17.0%	2.0%	2.0%	0.0%	81.0%
TG_8	Beatles Rubber	100	8.0%	9.0%	9.0%	1.0%	85.0%
TG_9	Kerry and Fonda	80	21.3%	13.8%	13.8%	0.0%	68.8%
TG_{10}	OJ Simpson	90	18.9%	2.2%	2.2%	0.0%	78.9%
	Average	93.5	23.5%	6.8%	6.8%	0.2%	72.7%

What's up with this near duplicate set?

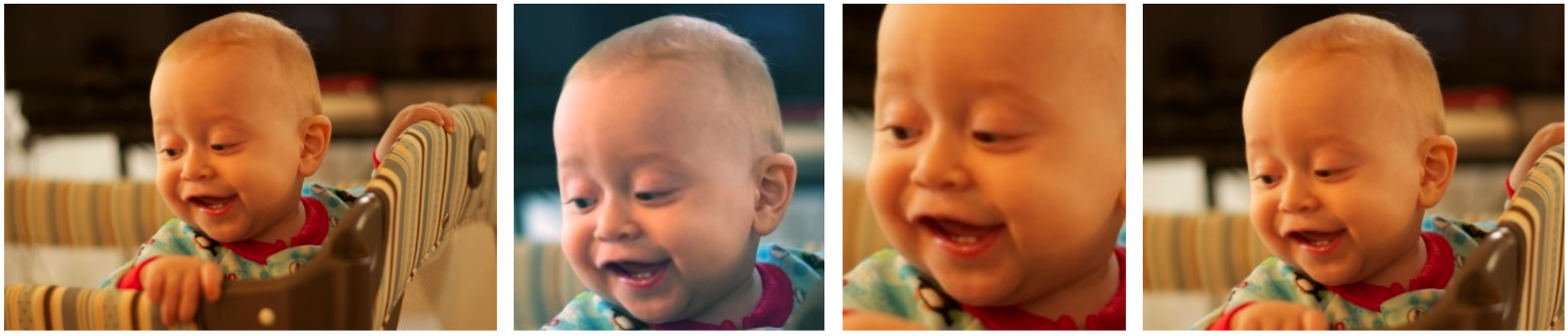


What's up with this near duplicate set?



What's up with this near duplicate set?

What's up with this near duplicate set?



Close-up

Close-up

First peek at Forests

- Forests (multiple co-existing trees) are a real case in real applications.
- Can our method be modified to find multiple trees?

Video Phylogeny Tree: VPT

Video Phylogeny Tree

- We ignore the sound track.
- We use only static image content.

Video Phylogeny Tree

- We ignore the sound track.
- We use only static image content.

Why not get one frame, and use the IPT as the VPT?

Video Phylogeny Tree

- We ignore the sound track.
- We use only static image content.

Why not get one frame, and
use the IPT as the VPT?

The IPTs of frames are different along the video!!!

Video Phylogeny Tree

- We ignore the sound track.
- We use only static image content.

Why not get one frame, and
use the IPT as the VPT?

The IPTs of frames are different along the video!!!

But why?













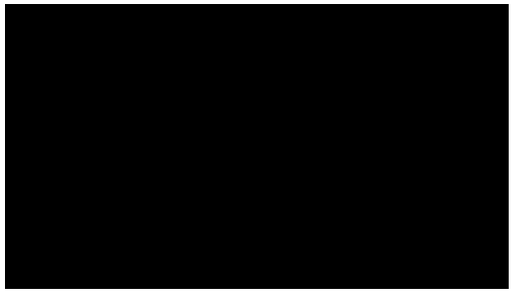


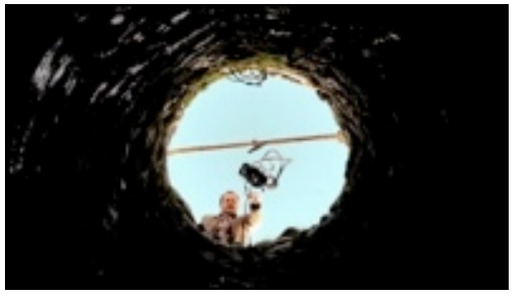


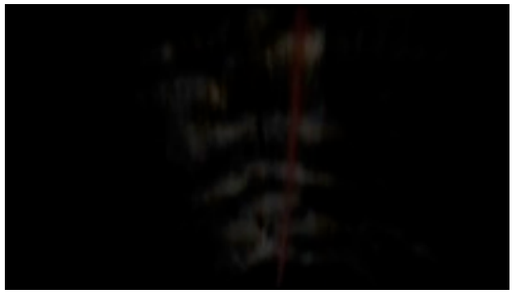


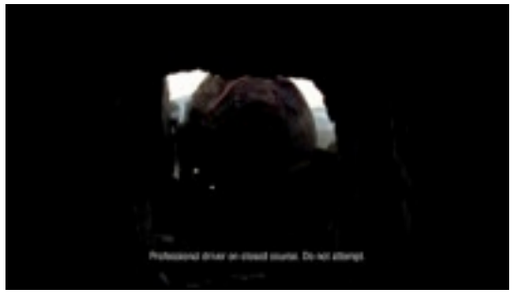
Professional driver on closed course. Do not attempt

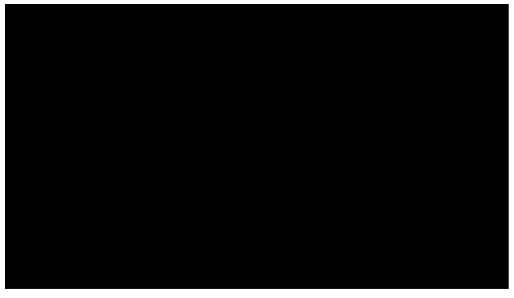
Professional driver on closed course. Do not attempt

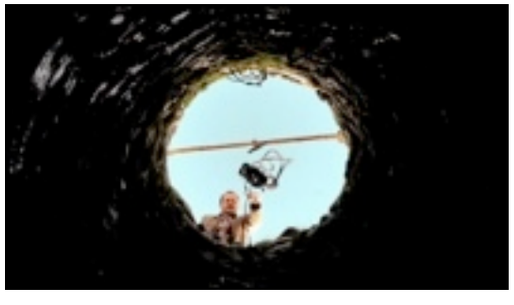


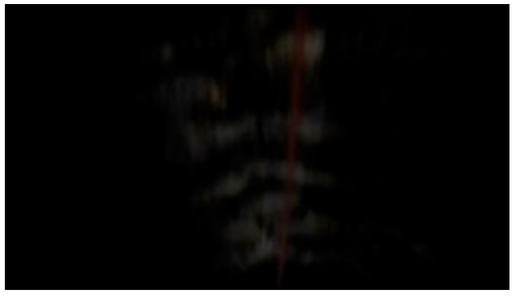


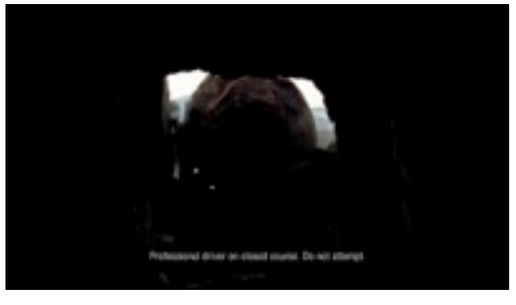


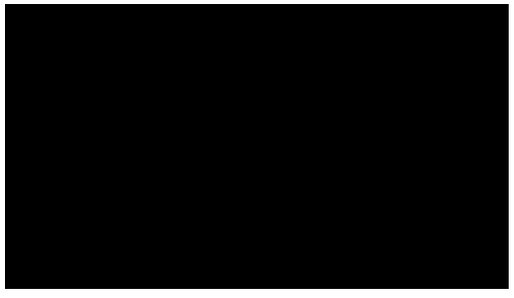




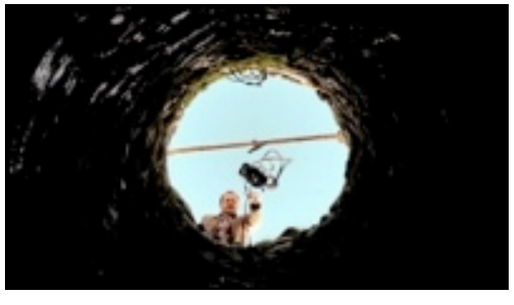
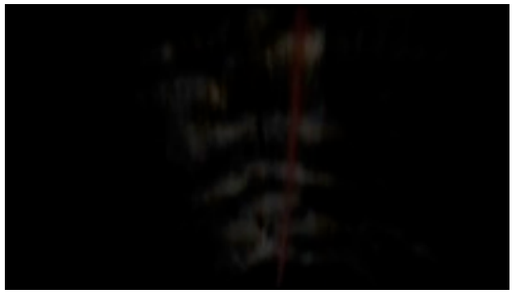


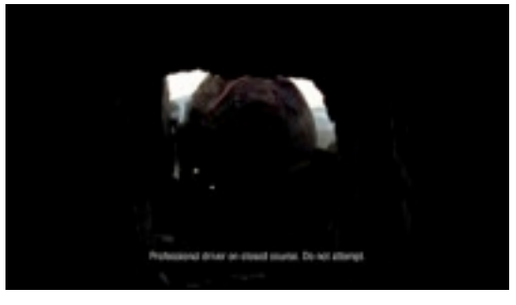


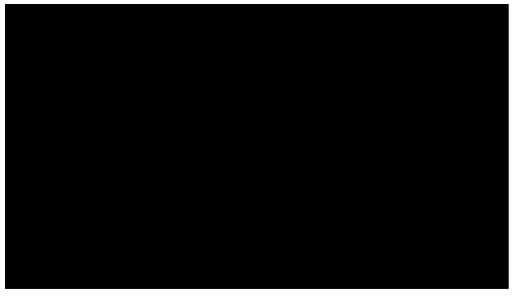




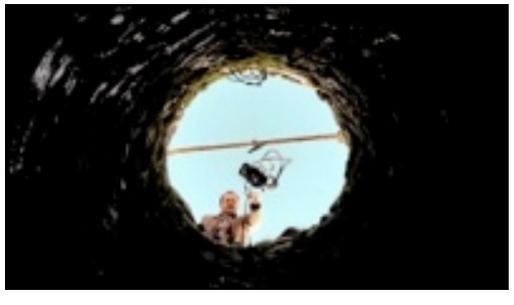
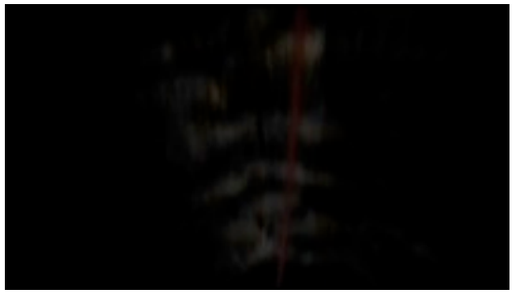


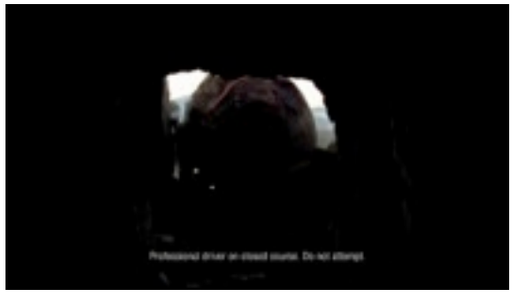





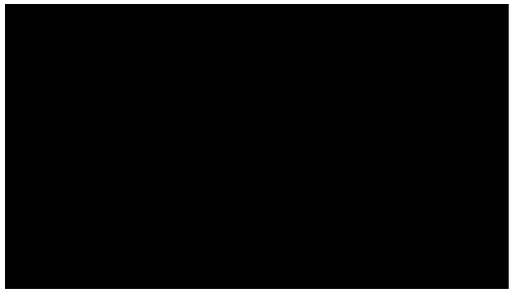





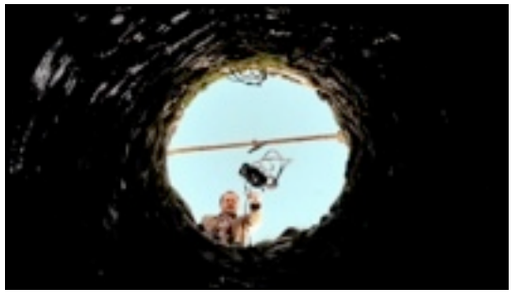
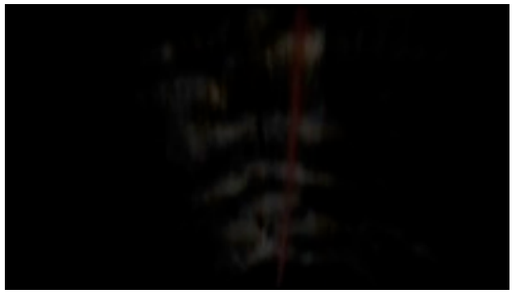



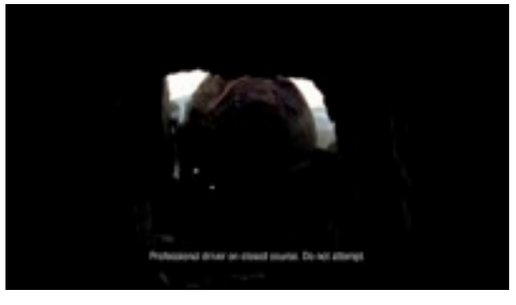




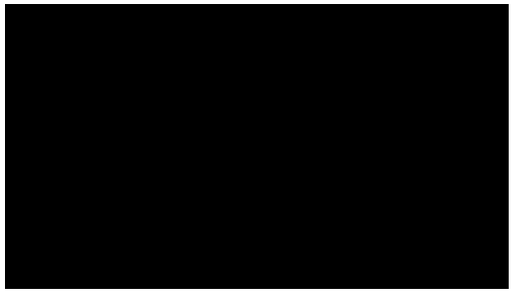




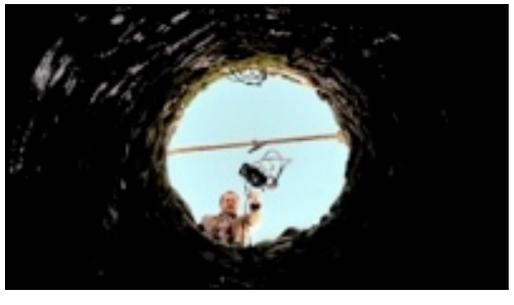
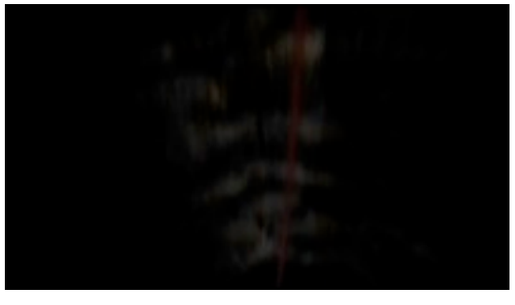



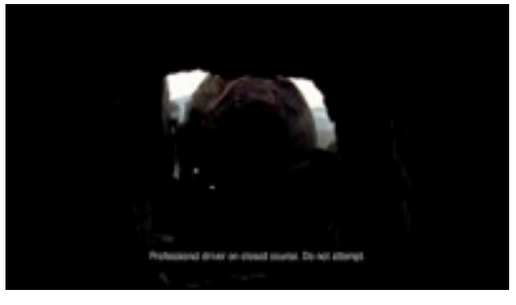




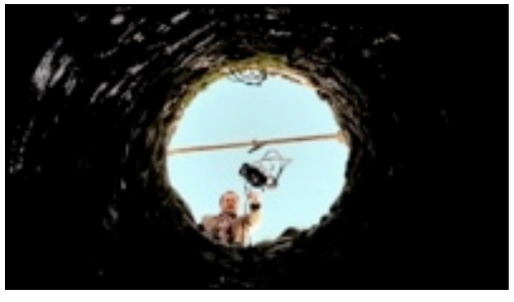


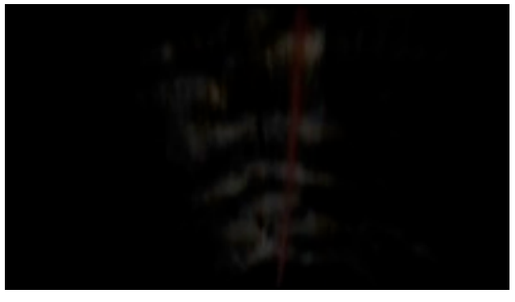



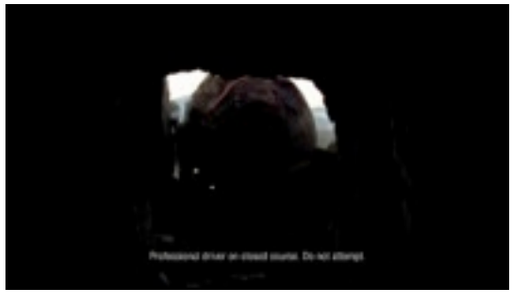
Professional driver on closed course. Do not attempt

Professional driver on closed course. Do not attempt



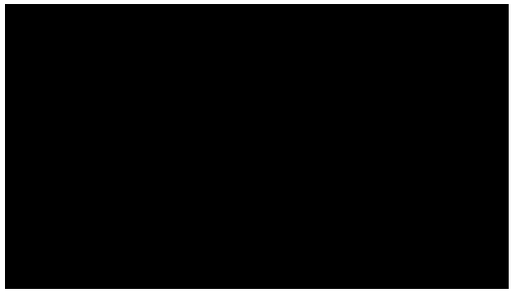


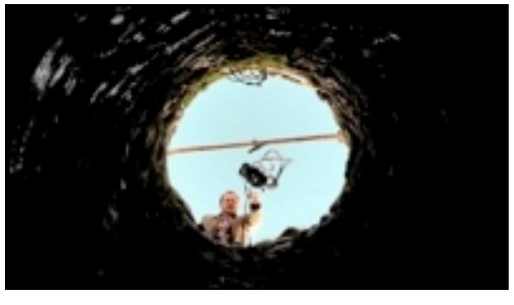


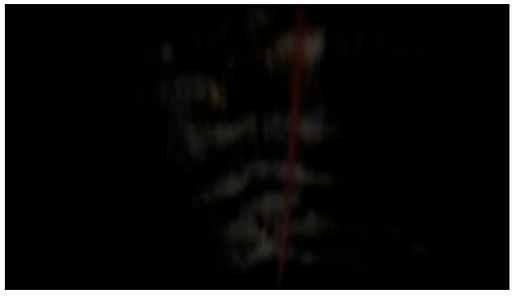


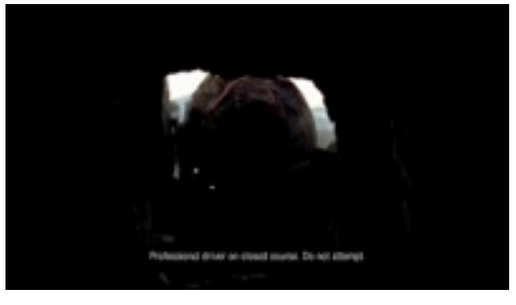


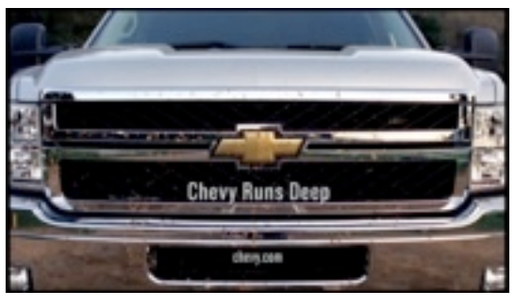

Chevy Runs Deep
chevy.com

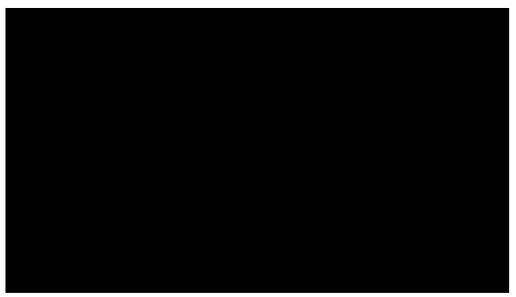









Chevy Runs Deep
chevy.com



Different IPTs

- Different quality over time, for example:
  - black frames,
  - blur,
  - compression artifacts,
  - dynamic range.

- Which (if any) is the right one?

A Few Approaches

A Few Approaches

- Expected result from a single frame IPT (baseline).

A Few Approaches

- Expected result from a single frame IPT (baseline).
- Minimum dissimilarity matrix followed by IPT.

A Few Approaches

- Expected result from a single frame IPT (baseline).
- Minimum dissimilarity matrix followed by IPT.
- Average dissimilarity matrix followed by IPT.

A Few Approaches

- Expected result from a single frame IPT (baseline).
- Minimum dissimilarity matrix followed by IPT.
- Average dissimilarity matrix followed by IPT.
- Reconciliation Tree.

Single-Frame Expectation

Single-Frame Expectation

- Calculate IPT on each frame.

Single-Frame Expectation

- Calculate IPT on each frame.

Single-Frame Expectation

- Calculate IPT on each frame.

- Calculate Expectation of metrics, but does not reconstruct a VPT (Video Phylogeny Tree).

Min / Average

Min / Average

- Sample frames.

Min / Average

- Sample frames.
- Calculate Dissimilarity Matrix on each synchronized frames.

Min / Average

- Sample frames.
- Calculate Dissimilarity Matrix on each synchronized frames.
- Create a new Dissimilarity Matrix using the frame's Dissimilarities
  - min,
  - average,
  - normalized min,
  - normalized average.

Min / Average

- Sample frames.
- Calculate Dissimilarity Matrix on each synchronized frames.
- Create a new Dissimilarity Matrix using the frame's Dissimilarities
  - min,
  - average,
  - normalized min,
  - normalized average.
- Construct VPT using oriented Kruskal on this new Matrix.

Reconciliation Approach

Reconciliation Approach

- Sample frames.

Reconciliation Approach

- Sample frames.
- Calculate IPT on each frame.

Reconciliation Approach

- Sample frames.
- Calculate IPT on each frame.
- Reconcile the frame's IPTs into the VPT.
  - Build Reconciliation Matrix.
  - Apply Tree Reconciliation Algorithm

Reconciliation Matrix

Algorithm 1 Reconciliation Matrix.

Require: number of near-duplicate videos, n
Require: number of selected frames, f
Require: 2-d vector, t, with the f phylogeny trees previously calculated

for $i \in [1..n]$ **do** ▷ Initialization
  for $j \in [1..n]$ **do**
    $P[i, j] \leftarrow 0$
  end for
end for
for $i \in [1..f]$ **do** ▷ Creating the matrix P
  for $j \in [1..n]$ **do**
    $P[j, t[i][j]] = P[j, t[i][j]] + 1$
  end for
end for
return P ▷ Returning the parenthood matrix P

Tree Reconciliation Alg.

Algorithm 2 Tree Reconciliation.

Require: number of near-duplicate videos, n
Require: matrix, P, from Algorithm 1

for $i \in [1..n]$ **do** ▷ Tree initialization
    $tree[i] \leftarrow i$
end for
$sorted \leftarrow$ sort positions (i, j) of P into nonincreasing order
▷ List of edges sorted from the most to the least common
$r \leftarrow 0$ ▷ Initially, the final root r is not defined
$n_{edges} \leftarrow 0$
for each position $(i, j) \in sorted$ **do** ▷ Testing each edge in order
    if $r = 0$ and $i = j$ **then** ▷ Defining the root of the tree
        $r \leftarrow i$
    end if
    if $i \neq r$ **then** ▷ If i is not the root of the tree
        if $\texttt{Root}(i) \neq \texttt{Root}(j)$ **then**
            if $\texttt{Root}(j) = j$ **then**
                $tree[j] \leftarrow i$
                $n_{edges} \leftarrow n_{edges} + 1$
                if $n_{edges} = n - 1$ **then** ▷ If the tree is complete
                    return $tree$ ▷ Returning the final VPT
                end if
            end if
        end if
    end if
end for

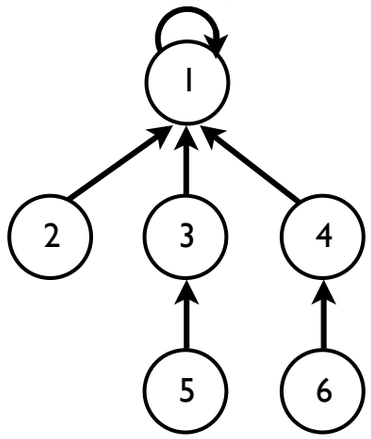
Reconciliation Approach

- Sample frames.
- Calculate IPT on each frame.
- Reconcile the frame's IPTs into the VPT.

- Is this enough to achieve good results?
- What are the limitations of this approach?

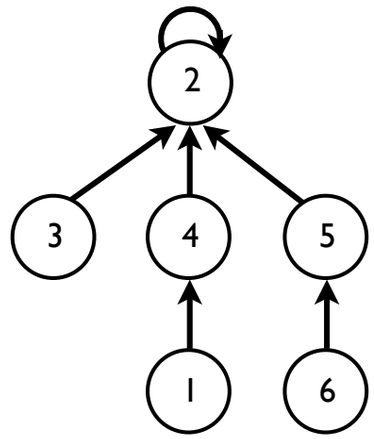
Example

Frame 1



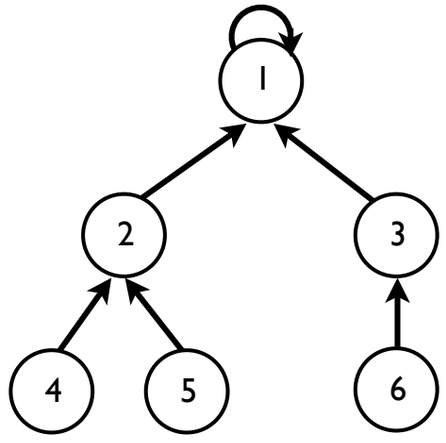
[1, 1, 1, 1, 3, 4]

Frame 2



[4, 2, 2, 2, 2, 5]

Frame 3



[1, 1, 1, 2, 2, 3]

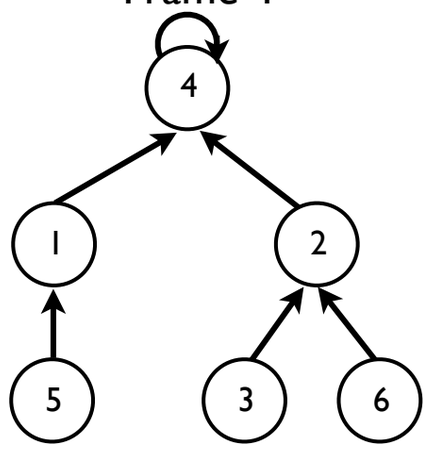
Parent

Child \ P	1	2	3	4	5	6
1	4			2		
2	3	1		2		
3	3	2				1
4	2	3		1		
5	1	3	2			
6	2	1	1	1	1	

Edges

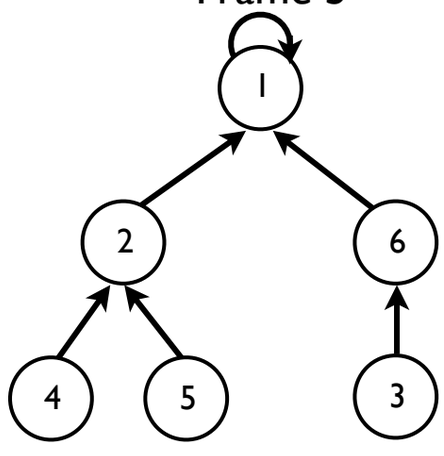
4	(1,1)	✔
3	(2,1)	✔
3	(3,1)	✔
3	(4,2)	✔
3	(5,2)	✔
2	(1,4)	✘
2	(2,4)	✘
2	(3,2)	✘
2	(4,1)	✘
2	(5,3)	✘
2	(6,1)	✔

Frame 4



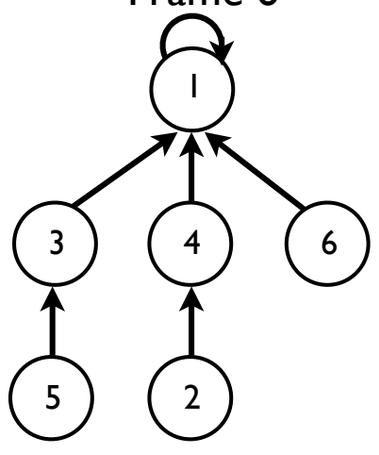
[4, 4, 2, 4, 1, 2]

Frame 5



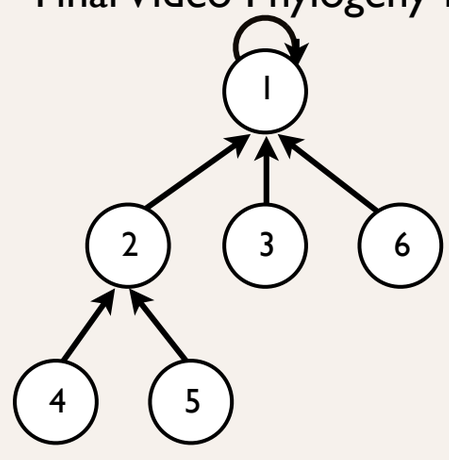
[1, 1, 6, 2, 2, 1]

Frame 6



[1, 4, 1, 1, 3, 1]

Final Video Phylogeny Tree



[1, 1, 1, 2, 2, 1]

Experimental results

Experimental results

- Limited experiments in this paper:
  - Ignored temporal cropping,
  - Ignored video compression on dissimilarities,
  - 16 Videos (Super Bowl Commercials 2011),
  - 16 trees,
  - 10 near-duplicates per tree.

Experimental results

- Limited experiments in this paper:
  - Ignored temporal cropping,
  - Ignored video compression on dissimilarities,
  - 16 Videos (Super Bowl Commercials 2011),
  - 16 trees,
  - 10 near-duplicates per tree.

- Transformations using `mencoder`.

Experimental results

- Limited experiments in this paper:
  - Ignored temporal cropping,
  - Ignored video compression on dissimilarities,
  - 16 Videos (Super Bowl Commercials 2011),
  - 16 trees,
  - 10 near-duplicates per tree.
- Transformations using `mencoder`.
- Sampling frames and sync by `ffmpeg`.

Experimental results

- Limited experiments in this paper:
  - Ignored temporal cropping,
  - Ignored video compression on dissimilarities,
  - 16 Videos (Super Bowl Commercials 2011),
  - 16 trees,
  - 10 near-duplicates per tree.
- Transformations using `mencoder`.
- Sampling frames and sync by `ffmpeg`.
- Dissimilarities using `OpenCV`.

Transformations

We used `mencoder` to generate the NearDuplicates, with these transformations and ranges:

Table I

TRANSFORMATIONS AND THEIR OPERATIONAL RANGES FOR CREATING THE CONTROLLED DATA SET.

Transformation	**Oper. Range**
(1) Global Resampling/Scaling (Up/Down)	$[90\%, 110\%]$
(2) Scaling by axis	$[90\%, 110\%]$
(3) Cropping	$[0\%, 5\%]$
(4) Brightness Adjustment	$[-10\%, 10\%]$
(5) Contrast Adjustment	$[-10\%, 10\%]$
(6) Gamma Correction	$[0.9, 1.1]$

Comparing Trees

Root: $R(\mathrm{IPT}_1, \mathrm{IPT}_2) = \begin{cases} 1, & \text{If } \texttt{Root}(\mathrm{IPT}_1) = \texttt{Root}(\mathrm{IPT}_2) \\ 0, & \text{Otherwise} \end{cases}$

Edges: $E(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|E_1 \cap E_2|}{n-1}$

Leaves: $L(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$

Ancestry: $A(\mathrm{IPT}_1, \mathrm{IPT}_2) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$

Results of Approaches

Table II
AVERAGE RESULTS FOR THE 16×16 TEST CASES UNDER CONSIDERATION FOR THE PROPOSED VPT METHODS.

Method	**Root**	**Depth**	**Edges**	**Leaves**	**Ancestry**
(E) Single Frame	76.5%	0.382	54.2%	67.7%	58.6%
(1) Min	59.0%	0.926	49.6%	64.1%	50.8%
(2) Min-Norm	68.0%	0.605	51.3%	66.4%	54.2%
(3) Avg	85.6%	0.215	56.6%	70.3%	62.0%
(4) Avg-Norm	85.9%	0.203	58.0%	72.4%	64.5%
(5) Reconc. Tree	91.0%	0.098	65.8%	77.7%	70.4%
(5)/(E) Boost	**18.9**%	**74.3**%	**21.4**%	**14.7**%	**20.1**%

Experimental Results

Table III
RESULTS FOR THE TREE RECONCILIATION APPROACH USING 16 DIFFERENT TREES.

Video	Root	Depth	Edges	Leaves	Ancestry
V_{01}	100.0%	0.000	68.1%	77.9%	73.4%
V_{02}	87.5%	0.125	66.9%	76.0%	68.8%
V_{03}	75.0%	0.312	56.9%	73.7%	57.8%
V_{04}	81.2%	0.188	57.5%	68.2%	60.7%
V_{05}	93.8%	0.062	69.4%	81.3%	73.9%
V_{06}	93.8%	0.125	66.2%	77.7%	72.7%
V_{07}	100.0%	0.000	73.1%	83.2%	79.5%
V_{08}	93.8%	0.062	59.4%	75.0%	66.1%
V_{09}	100.0%	0.000	70.6%	80.2%	73.1%
V_{10}	100.0%	0.000	65.6%	75.9%	72.3%
V_{11}	81.2%	0.188	64.4%	80.0%	69.8%
V_{12}	100.0%	0.000	68.7%	80.2%	76.4%
V_{13}	87.5%	0.125	75.0%	82.5%	77.7%
V_{14}	100.0%	0.000	69.4%	78.1%	72.5%
V_{15}	81.2%	0.188	56.9%	72.8%	64.2%
V_{16}	81.2%	0.188	65.0%	80.2%	67.2%
Average	91.0%	0.098	65.8%	77.7%	70.4%
Std Dev	8.8%	0.097	5.6%	4.0%	6.0%

Limitations of frame-based VPT

Limitations of frame-based VPT

- Does not use sound.

Limitations of frame-based VPT

- Does not use sound.
- Ignores temporal information of the visual content.

Limitations of frame-based VPT

- Does not use sound.
- Ignores temporal information of the visual content.
- Requires sync frames!
  - This is actually a very complicated issue in Video, and some codecs are finickier than others.
  - If we allow change in FPS + temporal crop, it might be impossible to fulfill this requisite.

Discussion

What's next?

What's next?

- Better study of the effect of the family of transformations on images and video.

What's next?

- Better study of the effect of the family of transformations on images and video.
- Use Reencoding for better video Dissimilarities.

What's next?

- Better study of the effect of the family of transformations on images and video.
- Use Reencoding for better video Dissimilarities.
- Global dissimilarity for video (instead of frame based).

What's next?

- Better study of the effect of the family of transformations on images and video.
- Use Reencoding for better video Dissimilarities.
- Global dissimilarity for video (instead of frame based).
- Reconstructing Forests.

What's next?

- Better study of the effect of the family of transformations on images and video.
- Use Reencoding for better video Dissimilarities.
- Global dissimilarity for video (instead of frame based).
- Reconstructing Forests.
- Large Scale Scenario.

Other Domains

Other Domains

- Software Development.
  - Phylogeny of code.

Other Domains

- Software Development.
  - Phylogeny of code.
- Computational Linguistics.
  - Phylogeny of Historical Documents.

Other Domains

- Software Development.
  - Phylogeny of code.
- Computational Linguistics.
  - Phylogeny of Historical Documents.
- File Carving through Phylogeny Algorithms.

Other Domains

- Software Development.
  - Phylogeny of code.
- Computational Linguistics.
  - Phylogeny of Historical Documents.
- File Carving through Phylogeny Algorithms.
- These cases are closer to the traditional Biological application, with metrics.

Other Domains

- Software Development.
  - Phylogeny of code.
- Computational Linguistics.
  - Phylogeny of Historical Documents.
- File Carving through Phylogeny Algorithms.
- These cases are closer to the traditional Biological application, with metrics.
- Beyond the Naïve Plagiarism Detection.

Thank you.