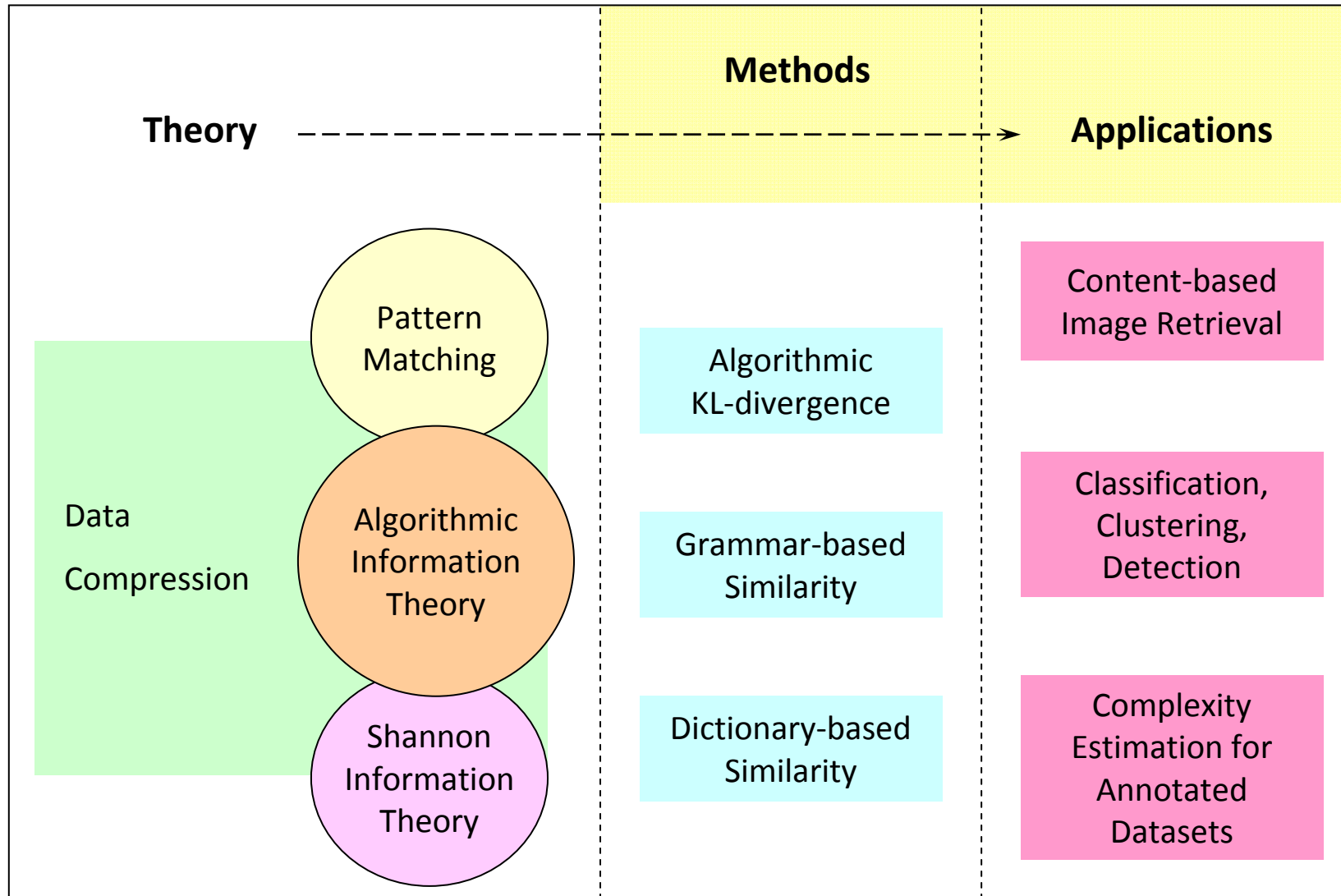


Pattern-oriented Algorithmic Complexity: Towards Compression-based Information Retrieval

PhD Thesis by Daniele Cerra

Director: Prof. Mihai Datcu

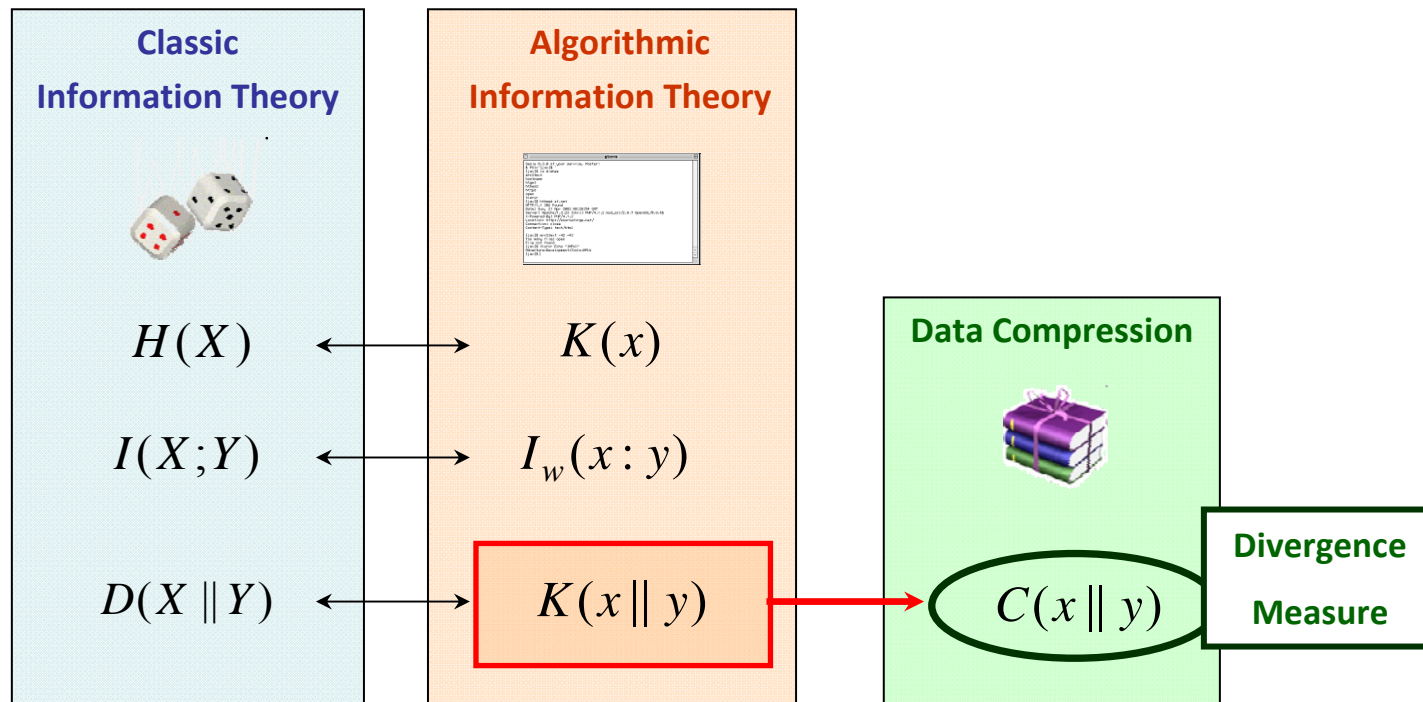


Main Contributions

Contributions (1/4)

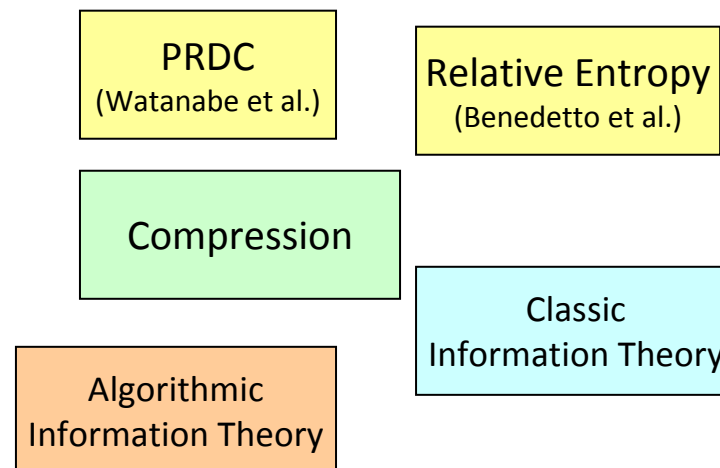
Competence Centre on Information Extraction and Image Understanding for Earth Observation

- Expansion Shannon/Kolmogorov correspondences
 - Definition of algorithmic relative complexity (or Kullback-Leibler divergence)
- Computable approximation based on data compression



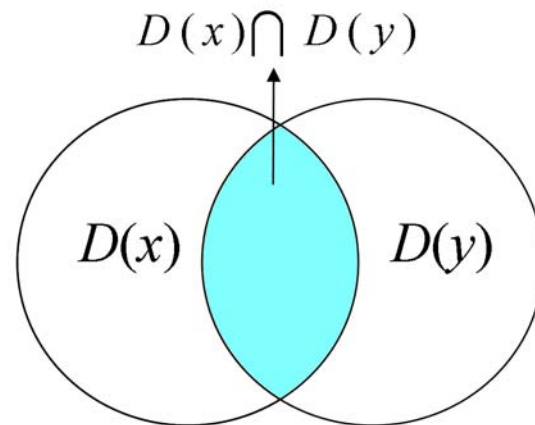
Contributions (2/4)

- Integrating the frame of compression-based methods
- Previously independently defined concepts are brought into the algorithmic information theoretical frame
 - Relative Entropy (RE), Benedetto et al., 2001
 - Pattern Representation based on Data Compression (PRDC), Watanabe et al., 2002



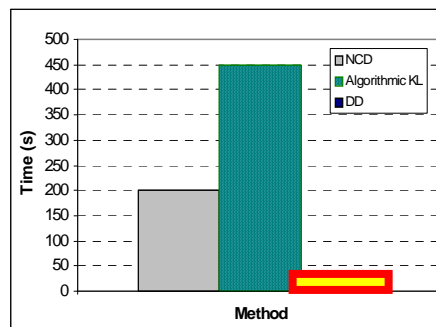
Contributions (3/4)

- Fast Compression Distance
- Reduced complexity with respect to previous techniques with no degradations in performance

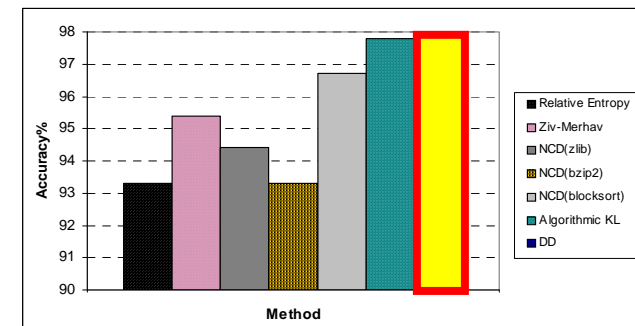


$$FCD(x, y) = \frac{|D(x)| - |\cap(D(x), D(y))|}{|D(x)|}$$

Computation Time



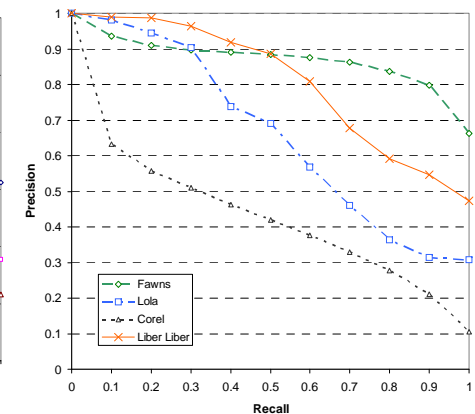
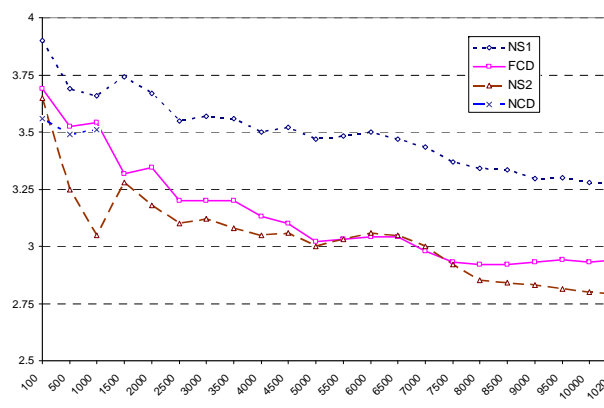
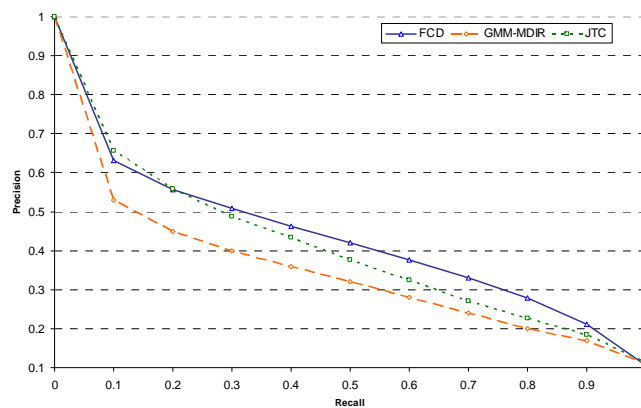
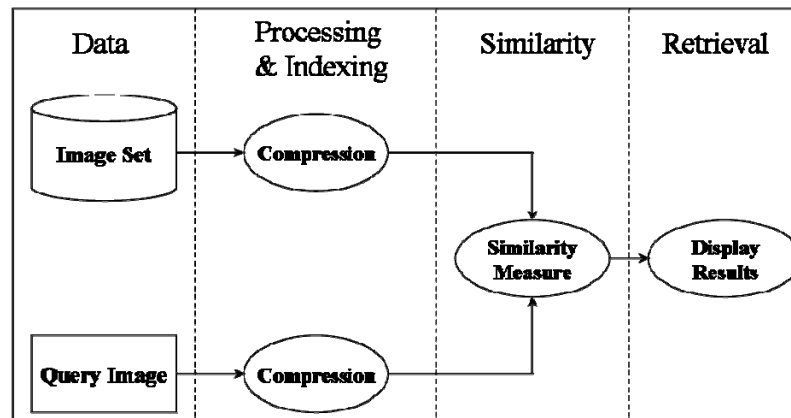
Accuracy



Contributions (4/4)

Competence Centre on Information Extraction and Image Understanding for Earth Observation

- Content Based Image Retrieval (CBIR) system based on data compression
- Experiments on datasets up to 100 times larger than in literature
- More thorough evaluation of compression-based similarity measures



Outline

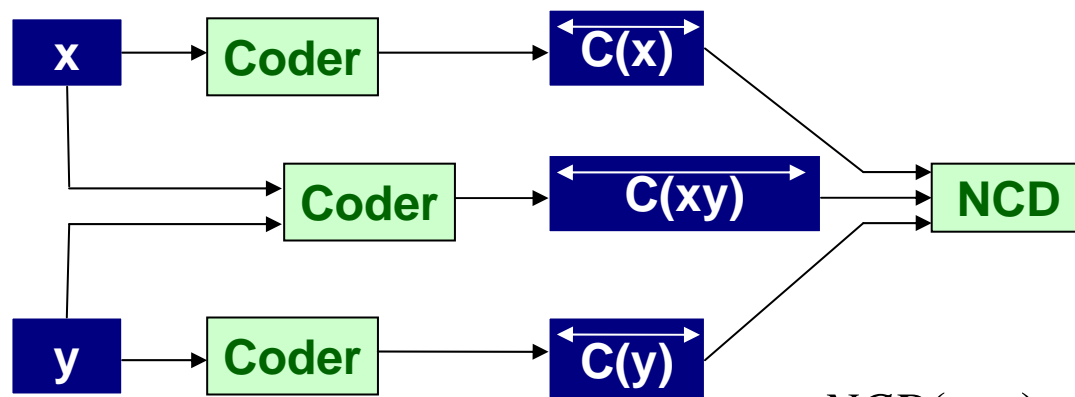
- The core: Compression-based similarity measures (CBSM)
- Theorectical Foundations
- Contributions: Theory
- Contributions: Applications and Experiments
- Conclusions and Perspectives

Outline

- **The core: Compression-based similarity measures (CBSM)**
- Theoretical Foundations
- Contributions: Theory
- Contributions: Applications and Experiments
- Conclusions and Perspectives

Compression-based Similarity Measures

- Most well-known: Normalized Compression Distance (NCD)
 - General Distance between any two strings x and y
Similarity metric under some assumptions
 - Basically parameter-free
 - Applicable with any off-the-shelf compressor (such as Gzip)
 - If two objects compress better together than separately, it means they share common patterns and are similar



$$NCD(x, y) = \frac{C(x, y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Evolution of CBSM

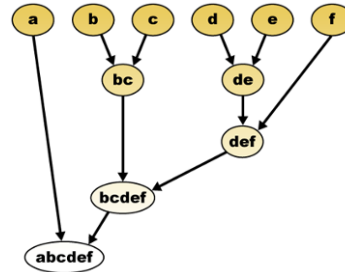
- 1993 Ziv & Merhav
 - First use of relative entropy to classify texts
- 2000 Frank et al., Khmelev
 - First compression-based experiments on text categorization
- 2001 Benedetto et al.
 - Intuitively defined compression-based relative entropy
 - Caused a rise of interest in compression-based methods
- 2002 Watanabe et al.
 - Pattern Representation based on Data Compression (PRDC)
 - Dictionary-based
 - First in classifying general data with a first step of conversion into strings
 - Independent from IT concepts
- 2004 NCD
 - Solid theoretical foundations (Algorithmic Information Theory)
- 2005-2006 Other similarity measures
 - Keogh et al. (Compression-based Dissimilarity Measure),
 - Chen & Li (Chen-Li Metric for DNA classification)
 - Sculley & Brodley (Cosine Similarity)
 - Differ from NCD only by their normalization factors - Sculley & Brodley (2006)
- 2008 Macedonas et al.
 - Independent definition of dictionary distance

Applications of CBSM

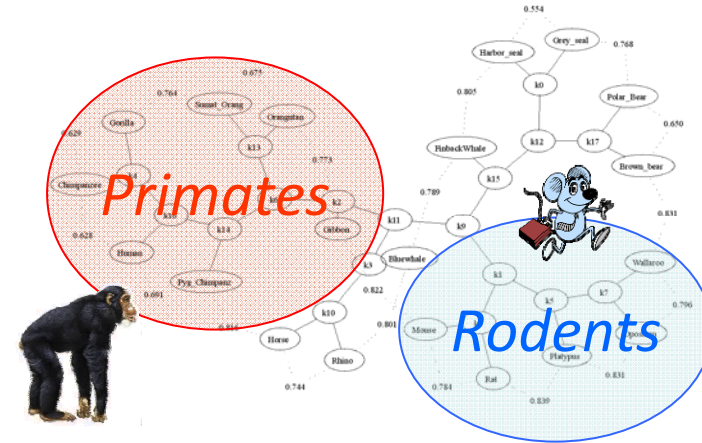
Competence Centre on Information Extraction and Image Understanding for Earth Observation

Clustering and classification of:

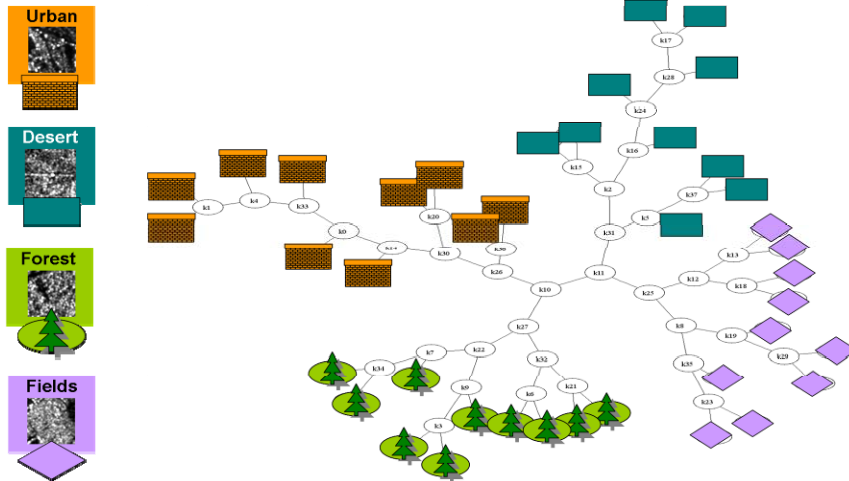
- Texts
- Music
- DNA genomes
- Chain letters
- Images
- Time Series
- ...



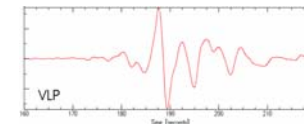
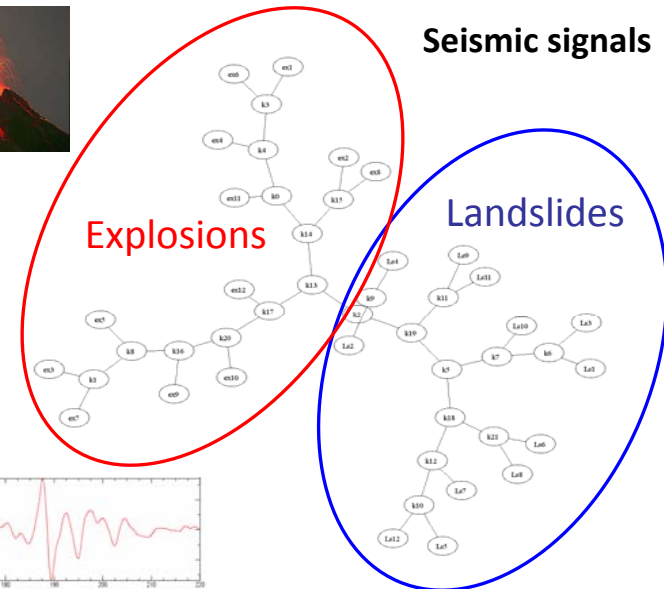
DNA Genomes



Satellite images



Seismic signals



Assessment and discussion of NCD results



- Results obtained by NCD often outperform state-of-the-art methods
 - Comparisons with 51 other distances*
- **But** NCD-like measures have always been applied to restricted datasets
 - Size < 100 objects in the main papers on the topic
 - All information retrieval systems use at least thousands of objects
 - More thorough experiments are required
- NCD is too slow to be applied on a large dataset
 - 1 second (on a 2.65 GHz machine) to process 10 strings of 10 KB each and output 5 distances
 - Being NCD data-driven, the full data has to be processed again and again to compute each distance from a given object
 - The price to pay for a parameter-free approach is that a compact representation of the data in any explicit parameter space is not allowed



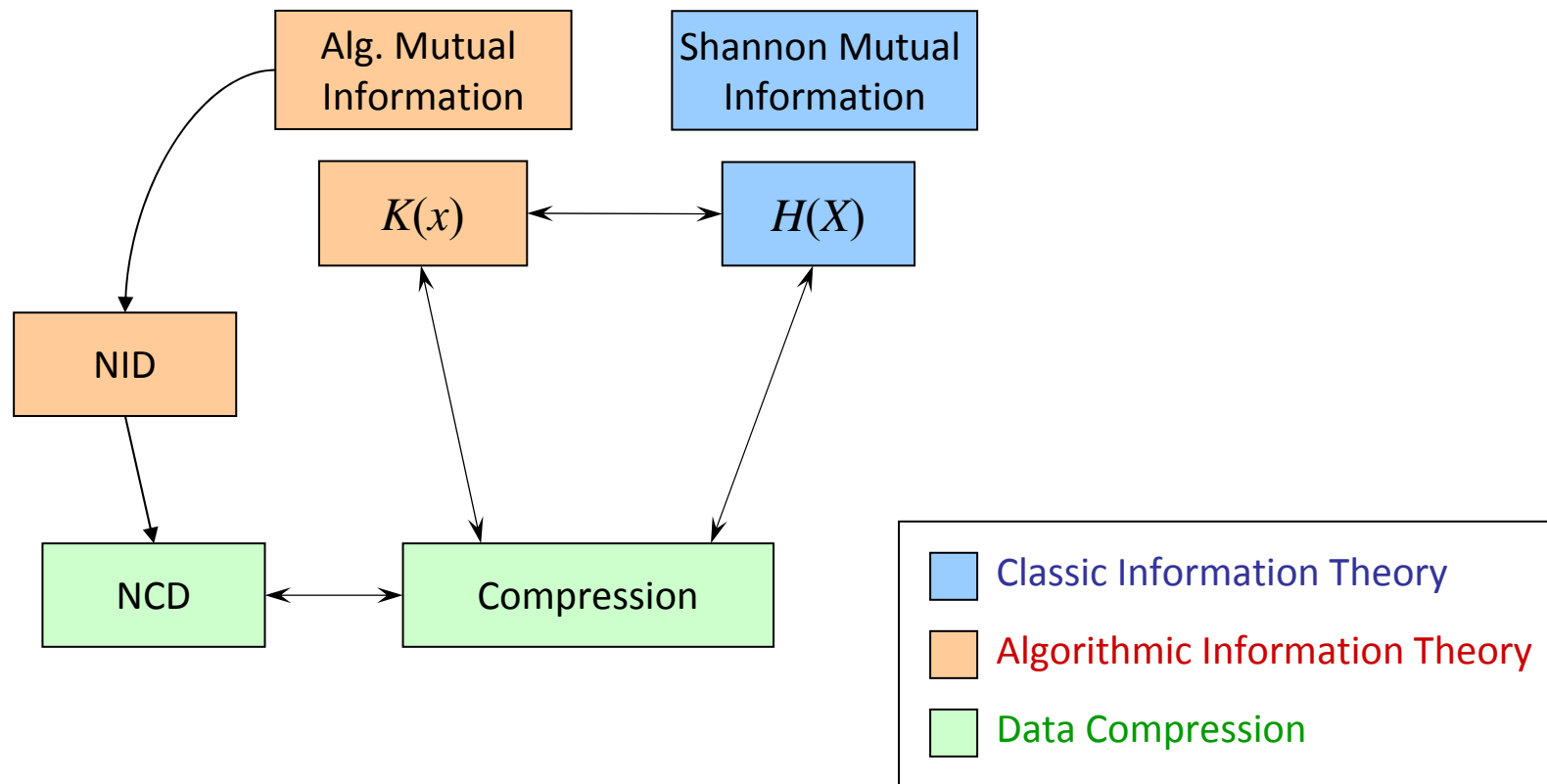
* Keogh, E., Lonardi, S. & Ratanamahatana, C. "Towards Parameter-free Data Mining", SIGKDD 2004.

Outline

- The core: Compression-based similarity measures (CBSM)
- **Theoretical Foundations**
- Contributions: Theory
- Contributions: Applications and Experiments
- Conclusions and Perspectives

A "Complex" Web

- How to quantify information?



Two approaches to information content

Probabilistic (classic)

VS.

Algorithmic

Information ↔ Uncertainty

Shannon Entropy

$$H(X) = - \sum_x p(x) \log p(x)$$

Information ↔ Complexity

Kolmogorov Complexity

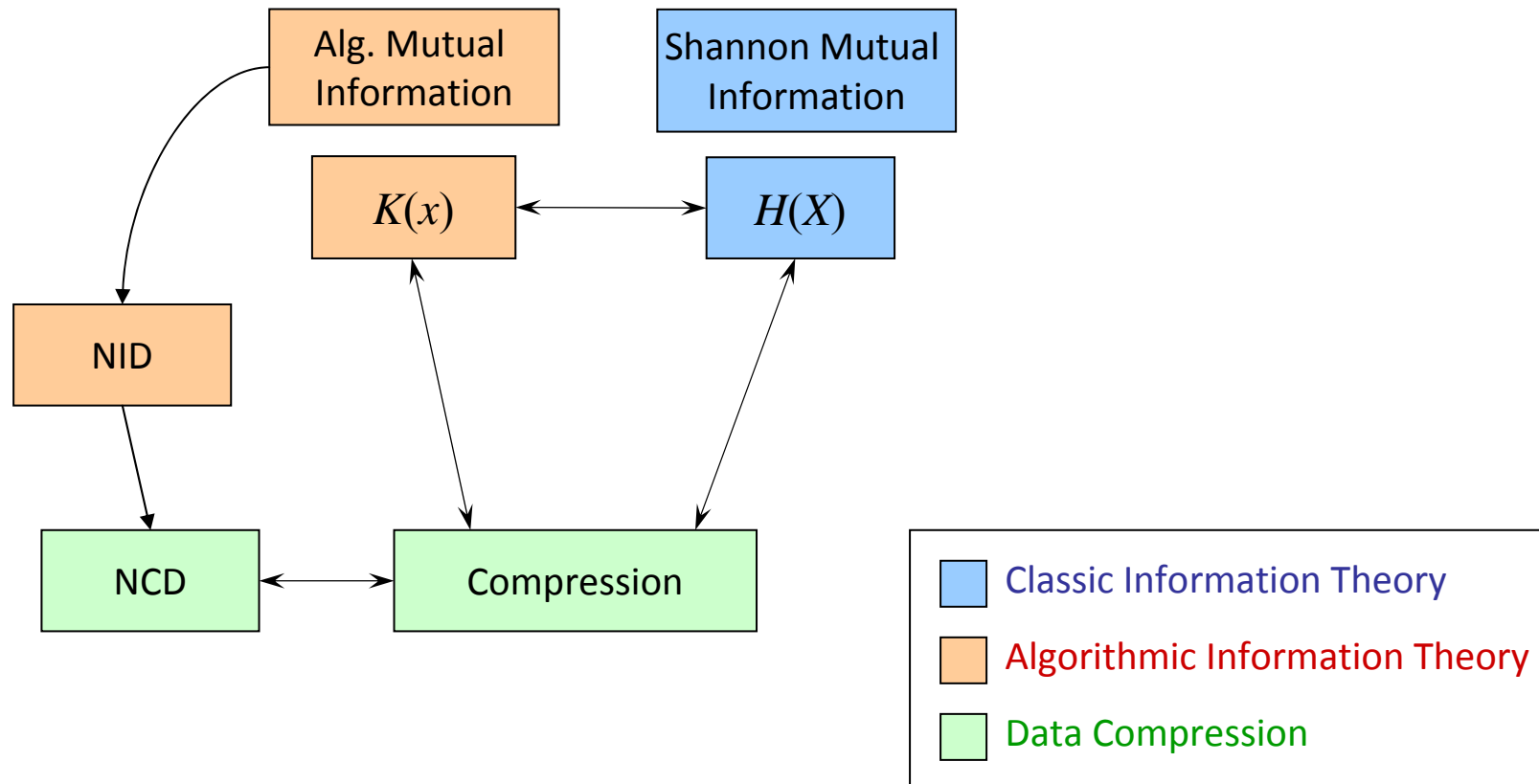
$$K(x) = \min_{q \in Q_x} |q|$$

- Related to a discrete **random variable** X on a finite alphabet A with a probability mass function $p(x)$
- Measure of the average uncertainty in X
- Measures the average number of bits required to describe X
- **Computable** if $p(x)$ is known

- Related to a **single object** x
- Length of the shortest program q among Q_x programs which outputs the finite binary string x and halts on a Universal Turing Machine
- Measures how difficult it is to describe x from scratch
- **Uncomputable**

A "Complex" Web

- How to measure the information shared between two objects?



Probabilistic (classic)

(Statistic) Mutual Information

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- Measure in bits of the amount of information a random variable X has about another variable Y
- The **joint entropy** $H(X,Y)$ is the entropy of the pair (X,Y) with a joint distribution $p(x,y)$
- Symmetric, non-negative
- If $I(X;Y) = 0$ then
 - $H(X;Y) = H(X) + H(Y)$
 - X and Y are **statistically independent**

VS.

Algorithmic

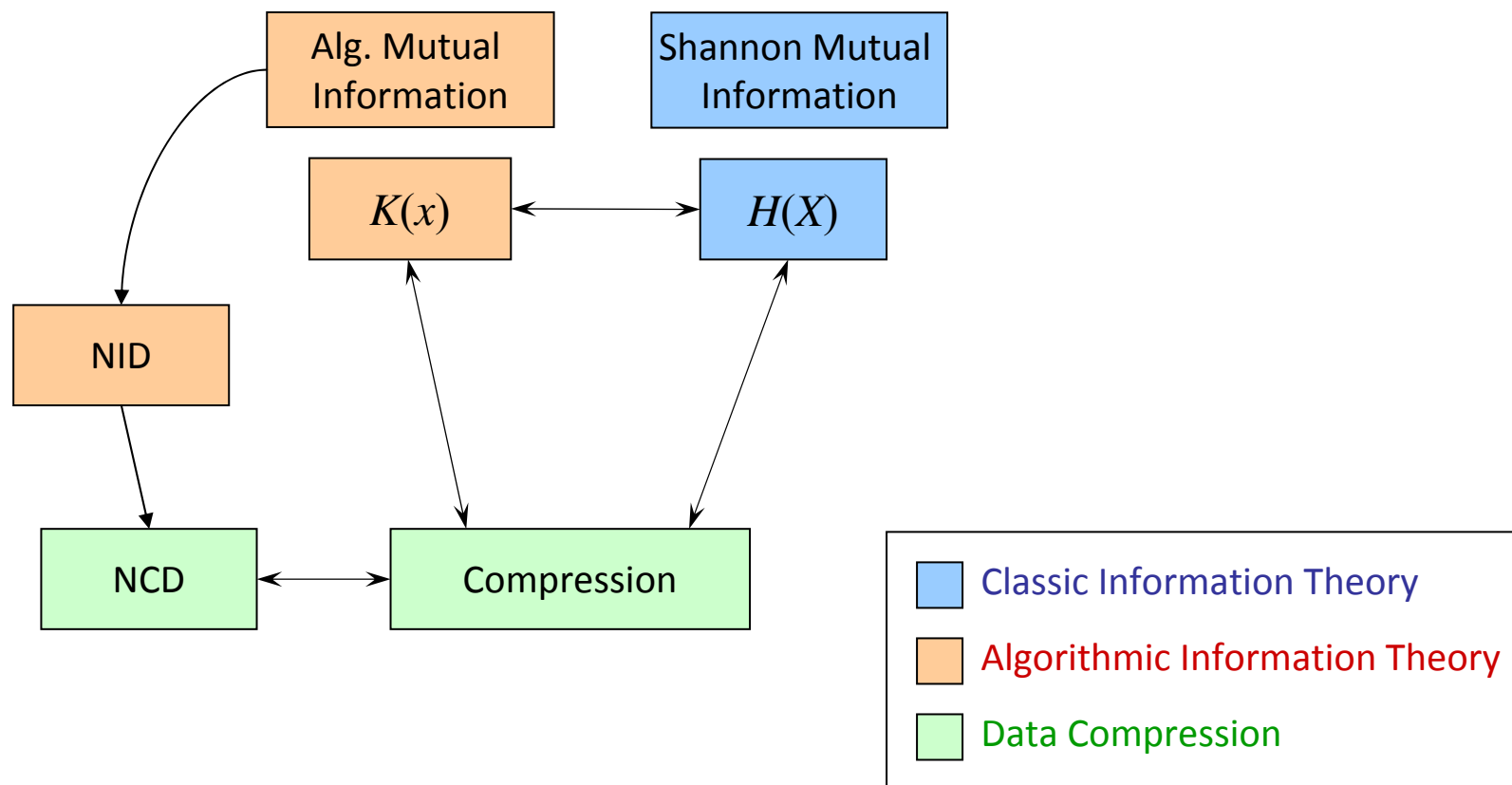
Algorithmic Mutual Information

$$I_w(x:y) = K(x) + K(y) - K(x,y)$$

- Amount of computational resources shared by the shortest programs which output the strings x and y
- The **joint Kolmogorov complexity** $K(x,y)$ is the length of the shortest program which outputs x followed by y
- Symmetric, non-negative
- If $I_w(x:y) = 0$ then
 - $K(x,y) = K(x) + K(y)$
 - x and y are **algorithmically independent**

A "Complex" Web

- How to derive a computable similarity measure?



Algorithmic

Normalized Information Distance (NID)

$$NID(x, y) = \frac{K(x, y) - \max\{K(x), K(y)\}}{\min\{K(x), K(y)\}}$$

$K(x) \rightarrow C(x)$
 \rightarrow

Computable

Normalized Compression Distance (NCD)

$$NCD(x, y) = \frac{C(x, y) - \max\{C(x), C(y)\}}{\min\{C(x), C(y)\}}$$

- Derived from algorithmic mutual information
- Normalized length of the shortest program that computes x knowing y , as well as computing y knowing x
- Similarity metric minimizing any admissible metric

- The size $K(x)$ of the shortest program which outputs x is assimilated to the size $C(x)$ of the compressed version of x
- Normalized measure of the elements that a compressor may use twice when compressing two objects x and y

Outline

- The core: Compression-based similarity measures (CBSM)
- Theorectical Foundations
- **Contributions: Theory**
 - **Expanding Shannon-Kolmogorov correspondences**
 - **Algorithmic relative complexity and computable approximation**
 - **Bringing Benedetto's relative entropy into the frame**
 - **Bringing Watanabe's PRDC into the frame**
 - **Considerations on Delta compression**
 - **Grammar-based distance**
- Contributions: Applications and Experiments
- Conclusions and Perspectives

Definition of relative complexity

Relative Entropy (Kullback-Leibler Divergence)

$$D(p_X \parallel p_Y) = \sum_i p_X(i) \log \frac{p_X(i)}{p_Y(i)}$$

$$= -\sum_i p_X(i) \log p_Y(i) + \sum_i p_X(i) \log p_X(i)$$

$$D(X \parallel Y) = H(X \oplus Y) - H(X)$$

- Measure of the distance between two probability mass functions p_X and p_Y related to two random variables X and Y
- Expected difference in the number of bits required to code an outcome i of X when using an optimal encoding for Y
- General case for mutual information
- $H(X \oplus Y)$ is the cross-entropy of X given Y

vs.

Relative Complexity

$$K(x \parallel y) = K(x \oplus y) - K(x)$$

- Difference between the computational resources needed to specify x in terms of its a description tailored for y , instead of its shortest description
- Compression power lost when compressing x by describing it only in terms of y , instead of using its most compact representation
- $K(x \oplus y)$ is the cross-complexity of x given y
- Algorithmic divergence measure

Definition of cross-complexity

$$K(x \parallel y) = K(x \oplus y) - K(x)$$

$$K(x \oplus y) = \min_{q_y \in Q_x} |q_y|$$

- The cross-complexity of x given y quantifies the computational resources needed by a Universal Turing Machine to specify x in terms of a description tailored for y
- The codes Q_x are forced to output x by a sequence of these operations:
 - Reuse part of the shortest code that outputs y
 - Use the command “Print s ”, where s is a substring of x
 - Any other way of compactly representing x is not allowed

Relative Entropy



Positively defined

$$D(X \parallel Y) \geq 0 \quad \forall X, Y$$

$$D(X \parallel Y) = 0 \quad \text{iff } X = Y$$

Asymmetric

$$\exists X, Y \mid D(X \parallel Y) \neq D(Y \parallel X)$$

Relative Complexity



Positively defined?

$$K(x \parallel y) \geq 0 \quad \forall x, y$$

$$K(x \parallel y) = 0 \quad \text{iff } x = y$$

Asymmetric?

$$\exists x, y \mid K(x \parallel y) \neq K(y \parallel x)$$

- Are the main properties of relative entropy maintained in the algorithmic frame?

Properties of Relative Complexity (1/2)

Lemma 1. The relative complexity of x related to y is positively defined

$$K(x \| y) \geq 0, \quad \forall x, y \quad \Leftrightarrow \quad K(x \oplus y) \geq K(x), \quad \forall x, y$$

$$K(x \oplus y) = K(x), \quad \text{if } x = y$$

- $K(x \oplus y)$ is a self-contained representation of x
- $K(x)$ is by definition the shortest representation of x
- Note that the stronger $K(x \oplus y) = K(x)$, *iff* $x = y$ does not hold

Lemma 2. The relative complexity of x (with known length $|x|$) related to y is upper bounded by $|x|$, plus an additive term

$$K(x \| y) = K(x \oplus y) - K(x) \quad K(x \oplus y) \leq |x|$$

- In the worst case no substring of x can be represented by the shortest code outputting y
- The shortest description is a command like “print the following $|x|$ bits: $x_0 x_1 x_2 \dots$ ”

$O(1)$

$|x|$

Properties of Relative Complexity (2/2)

Lemma 3. The relative complexity of x related to y is not symmetric

$$\exists x, y \mid K(x \parallel y) \neq K(y \parallel x)$$

- A and B : algorithmically independent finite binary strings of the same length
- Consider the strings x and y obtained by appending A to B and A to A
 - $x = \{A+B\}$
 - $y = \{A+A\}$
- Assume B is a simple sequence with respect to A such that $K(x) \cong K(y) \cong K(A)$
- $K(x \parallel y) - K(y \parallel x) = K(x \oplus y) - K(x) - K(y \oplus x) + K(y)$
- $\cong K(x \oplus y) - K(y \oplus x)$
- Note that y can be totally reconstructed by the optimal code to generate x , but the contrary is not true
- So $K(x \oplus y) > K(y \oplus x)$, and $K(x \parallel y) > K(y \parallel x)$

Relative Complexity Estimation

$$K(x \parallel y) = K(x \oplus y) - K(x)$$

$$C(x \parallel y) = C(x \oplus y) - C(x)$$

- Pseudocode to compute a “cross compression” $C(x \oplus y)$:
 - Assuming to have available a set of n explicit dictionaries $Dic(y,p)$, each containing the substrings found within a string y of length n until each position $p=0..n-1$ in y

1. Position $p=0$.
2. If $p = |x|$, then Halt.
3. Consider the symbol x_p at position p . If the partial dictionary $Dic(y,p)$ contains a word starting with x_p , then:
 - a. Output the code of a pattern c of length n contained in $Dic(y,p)$ matching a substring of x starting at x_p , chosen so that n is maximal and $p+n \leq |x|$
 - b. $p=p+n$
 - c. Go to 2
4. Output x_p .
5. $p=p+1$
6. Go to 2

Normalized Relative Complexity



$$K(x \parallel y) = K(x \oplus y) - K(x)$$

$$C(x \parallel y) = C(x \oplus y) - C(x)$$

$$\bar{C}(x \parallel y) = \frac{C(x \oplus y) - C(x)}{|x| - C(x)}$$

- Computable between any two strings x and y
- Estimates the effectiveness in compressing x when a parallel processing of y is simulated, with the compressor only learning the model of y
- Results in a similarity measure ranging from 0 (total similarity) to 1

Benedetto's Relative Entropy

- Benedetto et al. "relative entropy" (2001) $H_r(x || y)$
 - Append to a file y a small fraction Δx of x
 - Compress $y + \Delta x$
 - Assume that Δx is compressed by the model learned from y

The terms can be matched with the terms of normalized relative complexity

$$\bar{C}(x || y) = \frac{C(x \oplus y) - C(x)}{|x| - C(x)}$$

$$H_r(x || y) = \frac{C(y + \Delta x) - C(y) - (C(x + \Delta x) - C(x))}{|\Delta x|}$$

Compression of x with model of y

Compression with model of y conditioned by Δx model

The entire string x can be analyzed

Restricted to a small fraction Δx of x

Distance ranges from 0 to 1

Normalization term makes the distance always < 1 ,
if Δx is compressible




Authorship Attribution

- We assume that the relative complexity $C(x||y)$ should perform better than $H_r(x || y)$
- We want to recognize the author of an unknown text
- A corpus of 90 texts of known Italian authors is given
- Use each text t as query and assign it to the author of the most similar text retrieved s
 - The relative complexity $C(t||s)$ is minimal

| Author | Dante | D'Annunzio | Deledda | Fogazzaro | Guicciardini | Machiavelli | Manzoni | Pirandello | Salgari | Svevo | Verga | TOT |
|------------------|----------|------------|-----------|-----------|--------------|-------------|----------|------------|-----------|----------|----------|-----------|
| Texts | 8 | 4 | 15 | 5 | 6 | 12 | 4 | 11 | 11 | 5 | 9 | 90 |
| Successes | 8 | 4 | 15 | 3 | 6 | 12 | 4 | 11 | 11 | 5 | 9 | 88 |

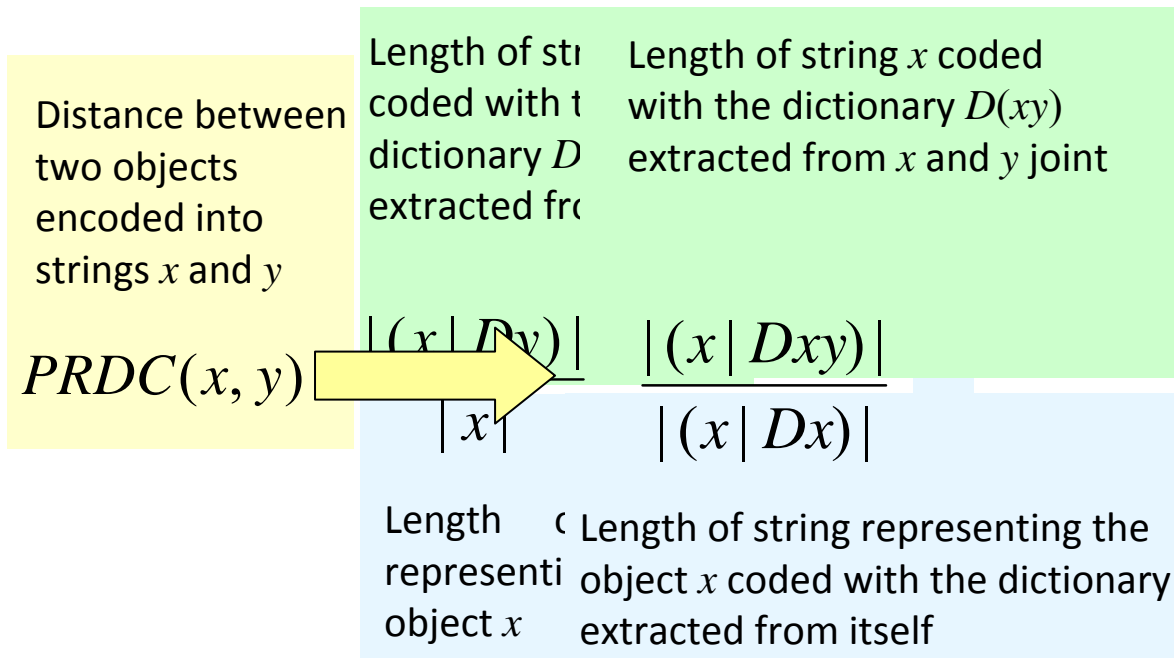
- Accuracy
 - Relative Complexity: 97.8 %
 - Relative entropy (same dataset): 93.3%
- The concept of relative entropy is now inserted in a new theoretical frame and can be better understood

Some Considerations on Delta Compression

| | |
|---|---|
|  | x $C(x)$ 135 KB |
|  | y $C(y)$ 139 KB |
|  | $\Delta(x, y)$ $C(x y)$ 22 KB |

- Delta compression $\Delta(x, y)$ represents a target file x with respect to a source file y .
- The Delta file $\Delta(x, y)$ is as compact as possible and contains the information to fully recover x if y is available.
- The conditional Kolmogorov complexity $K(x | y)$ is defined as the shortest program which outputs x if y is given “for free” as an auxiliary input for the computation.
- $\Delta(x, y)$ can be regarded as a way to estimate $K(x | y)$ through a conditional compression $C(x | y)$.

Watanabe et al., 2002



- PRDC equation is not normalized according to the complexity of x and skips the joint compression step.
- Normalizing the equation used in PRDC, almost identical measure with NCD are obtained ($O(10^{-4})$ average difference on 400 measures)
- PRDC can be inserted in the list of measures which differ by NCD only for the normalization factor (Sculley & Brodley, 2006)

Grammar-based Approximation

- A dictionary extracted from a string x in PRDC may be regarded as a model for x
- To better approximate $K(x)$, consider the smallest Context-Free Grammar (CFG) generating x .
 - The grammar's set of rules can be regarded as the smallest dictionary and generative model for x .

Sample CFG $G(z)$ for string
 $z = \{aaabaaacaaadaaaf\}$

$A \rightarrow aaa$
 $S \rightarrow AbAcAdAe$

$$K(x) \rightarrow Cg(x) = \begin{cases} N, & \text{if } N \leq 1 \\ C_x + \left(1 - \frac{\log_2 N}{\log_2 C_x + |G(x)|}\right) |G(x)|, & \text{o.w.} \end{cases}$$

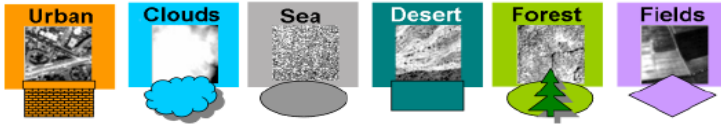
C_x : size of x of length N represented by its smallest context-free grammar $G(x)$
 $|G(x)|$: number of rules contained in the grammar

- Two-part complexity representation
 - Model + data given the model (MDL-like)
 - Complexity overestimations are intuitively accounted for and decreased in the second term

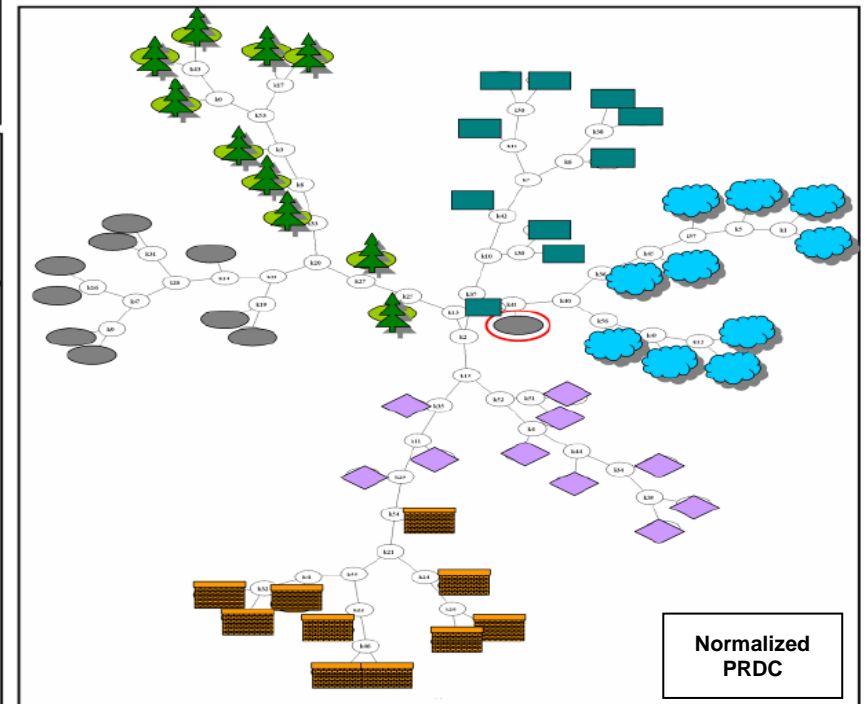
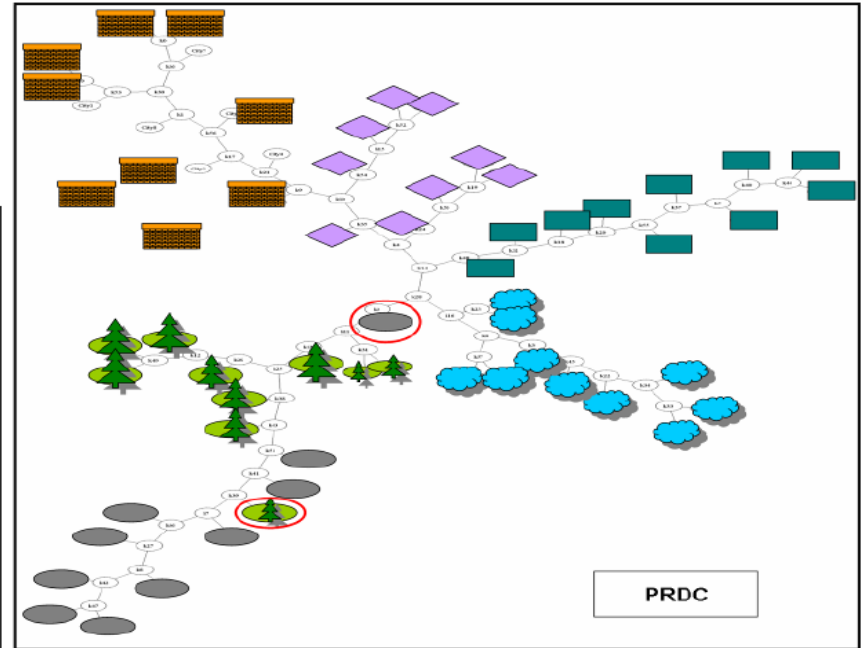
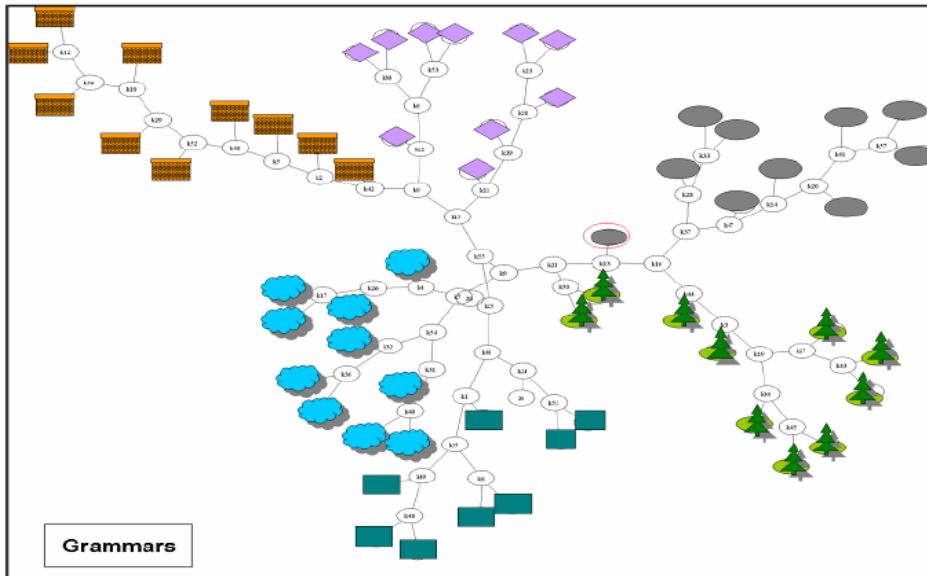
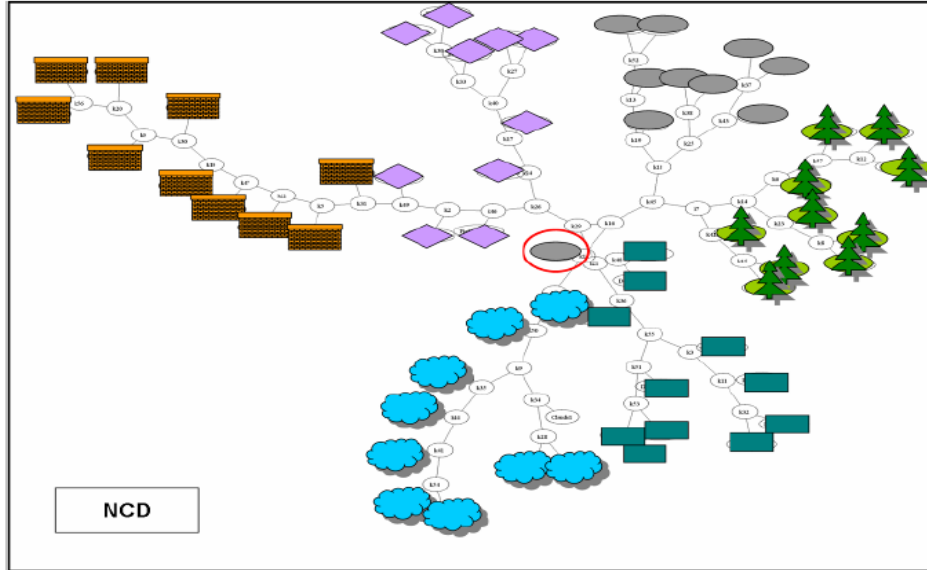
Avg. distances on 40,000 measurements

| NCD with.. | Intraclass | Interclass | Difference |
|-----------------------|------------|------------|------------|
| Standard Compressor | 1.02 | 1.11 | 0.09 |
| Grammar approximation | 0.83 | 0.96 | 0.13 |

Comparisons...



Competence Centre on Information Extraction and Image Understanding for Earth Observation



Drawbacks of the Introduced CBSM

| Similarity Measure | Accuracy | Speed |
|------------------------------|-------------------|-------------------|
| NCD | Comparable to NCD | Comparable to NCD |
| Algorithmic Kullback-Leibler | Comparable to NCD | Worse than NCD |
| PRDC | Worse than NCD | Better than NCD |
| Normalized PRDC | Comparable to NCD | Comparable to NCD |
| Grammar-based | Better than NCD | Worse than NCD |
| ?? | Better than NCD | Better than NCD |

| | |
|--|-------------------|
| | Better than NCD |
| | Comparable to NCD |
| | Worse than NCD |

| |
|--|
| Solution |
| <ul style="list-style-type: none"> ▪ Extract dictionaries from the data, possibly offline ▪ Compare only those |

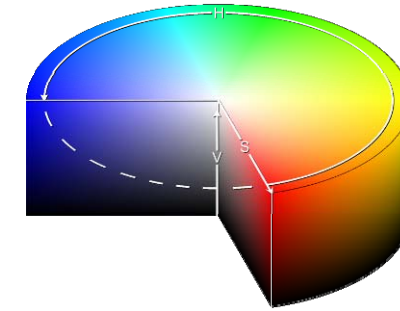
How to combine accuracy and speed?

Outline

- The core: Compression-based similarity measures (CBSM)
- Theorectical Foundations
- Contributions: Theory
- **Contributions: Applications and Experiments**
 - **Fast Compression Distance**
 - **Parameter free Content-based Image Retrieval System**
 - **Experiments on Earth Observation data**
- Conclusions and Perspectives

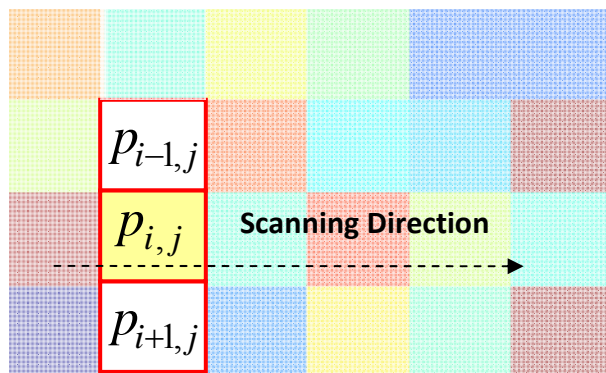
1D encoding for images

- Conversion to Hue Saturation Value (HSV) color space
- Scalar quantization
 - 4 bits for Hue
 - Human eye is more sensitive to changes in hue
 - 2 bits for Saturation
 - 2 bits for Value



HSV color space

- What about loss of textural information?
- Horizontal textural information is already implicit in the dictionaries
- Basic vertical interactions are stored for each pixel
 - Smooth / Rough: 1 bit of information
- Other solutions (e.g. Peano scanning) gave worse performances



$$v(p_{i,j}) = \begin{cases} 1, & \text{if } (d(p_{i,j}, p_{i+1,j}) > t) \parallel (d(p_{i,j}, p_{i-1,j}) > t) \\ 0, & \text{otherwise} \end{cases}$$

Dictionary-based Distance: Dictionary Extraction



LZW

- Dictionary-based universal compression algorithm
- Improvement by Welch (1984) over the LZ78 compressor (Lempel & Ziv, 1978)
- Searches for matches between the text to be compressed and a set of previously found strings contained in a dictionary
- When a match is found, a substring is substituted by a code representing a pattern in the dictionary

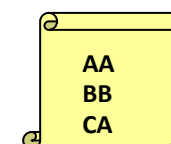
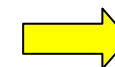
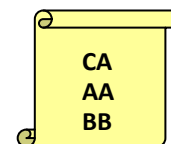
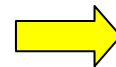
”TOBEORNOTTOBEORTOBEORNOT!”

| Current Char | Next Char | Output | Added to Dictionary |
|--------------|-----------|--------|---------------------|
| Null | T | | |
| T | O | T | TO=< 27 > |
| O | B | O | OB=< 28 > |
| B | E | B | BE=< 29 > |
| E | O | E | EO=< 30 > |
| O | R | O | OR=< 31 > |
| R | N | R | RN=< 32 > |
| N | O | N | NO=< 33 > |
| O | T | O | OT=< 34 > |
| T | T | T | TT=< 35 > |
| TO | B | < 27 > | TOB=< 36 > |
| BE | O | < 29 > | BEO=< 37 > |
| OR | T | < 31 > | ORT=< 38 > |
| TOB | E | < 36 > | TOBE=< 39 > |
| EO | R | < 30 > | EOR=< 40 > |
| RN | O | < 32 > | RNO=< 41 > |
| OT | ! | < 34 > | OT! |

- Convert each image to a string and extract meaningful patterns into dictionaries
 - Unlike LZW, loose (or no) constraints on dictionary size, flexible alphabet size
- Sort entries in the dictionaries in order to enable binary searches
- Store only the dictionary



..ABABBCA..



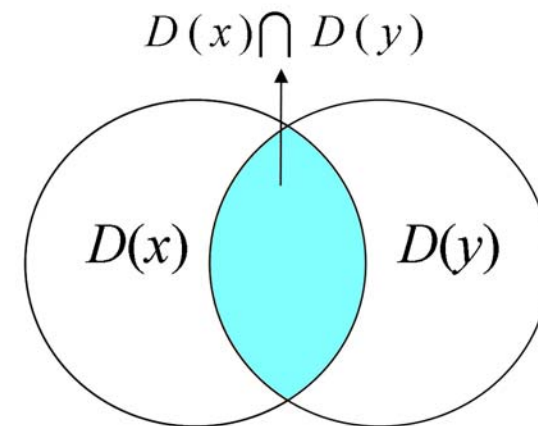
Fast Compression Distance

- Consider two dictionaries to compute a distance between them as the difference between shared/not shared patterns
- The joint compression step of NCD is now replaced by an inner join of two sets
- NCD acts like a black box, FCD simplifies it by making dictionaries explicit
 - Dictionaries have to be extracted only once (also offline)

```
Count(select * from (D(x)))
```

$$FCD(x, y) = \frac{|D(x)| - |\cap(D(x), D(y))|}{|D(x)|}$$

```
Count(select * from Inner_Join(D(x), D(y)))
```



How fast is FCD with respect to NCD?

Operations needed for the joint compression steps (LZW-based NCD)

| | | |
|------------------------------------|-------------------------------|--------------------------|
| $FCD(x, y) = \bigcap (D(x), D(y))$ | $m_x \log m_y$ | n_x n. elements in x |
| $NCD(x, y) = C(x, y)$ | $(n_x + n_y) \log(m_x + m_y)$ | m_x n. patterns in x |

- Further advantages
 - If in the search a pattern gives a mismatch, ignore all extensions of that pattern
Ensured by LZW's prefix-closure property
 - Ignore shortest patterns (regard them as noise)
 - To reduce storage space, ignore all redundant patterns which are prefixes of others
No losses also ensured by LZW's prefix-closure property
- Complexity decreases by approx. one order of magnitude

Datasets



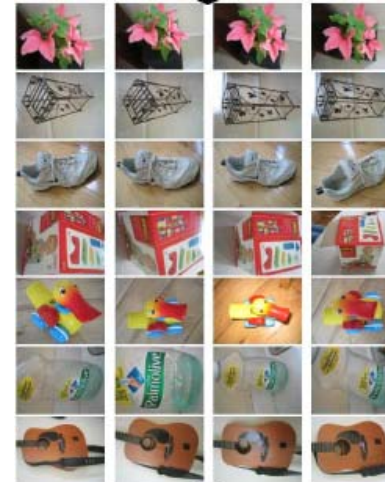
Corel
1500 digital photos and hand-drawn images



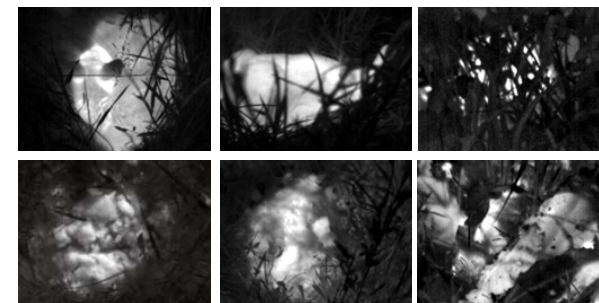
Lola
164 video frames from the movie "Run, Lola, Run"



Liber Liber
90 books of known Italian authors



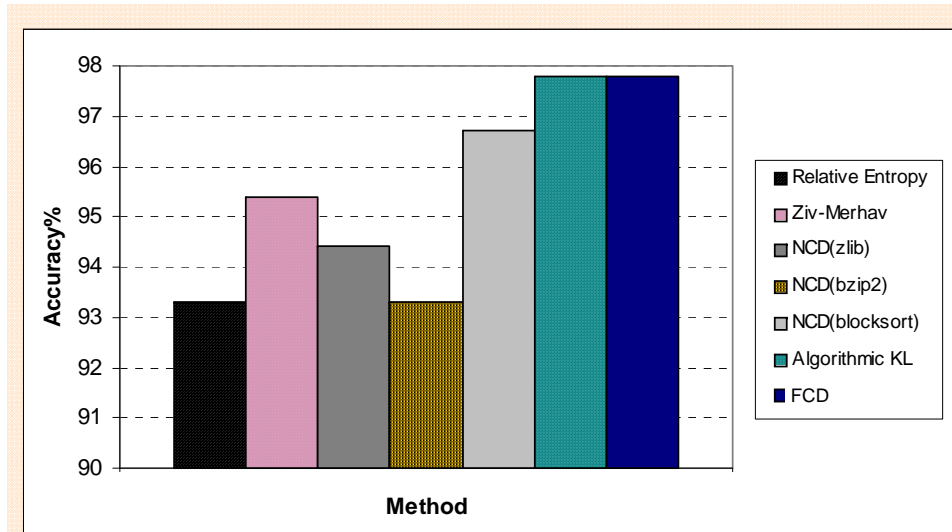
Nister-Stewenius
10,200 photographs of objects pictured from 4 different points of view



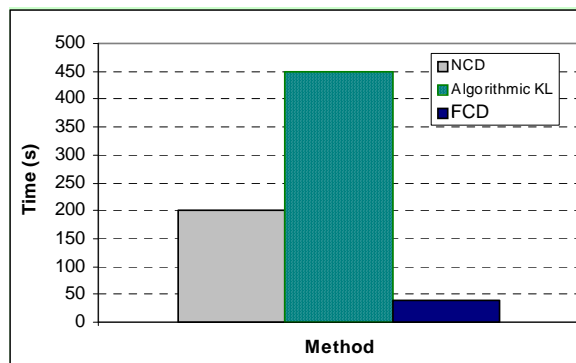
Fawns & Meadows
144 infrared images of meadows some of which contain fawns

Authorship Attribution

| Author | Texts | Successes |
|--------------|-----------|-----------|
| Dante | 8 | 8 |
| D'Annunzio | 4 | 4 |
| Deledda | 15 | 15 |
| Fogazzaro | 5 | 5 |
| Guicciardini | 6 | 6 |
| Machiavelli | 12 | 10 |
| Manzoni | 4 | 4 |
| Pirandello | 11 | 11 |
| Salgari | 11 | 11 |
| Svevo | 5 | 5 |
| Verga | 9 | 9 |
| TOTAL | 90 | 88 |

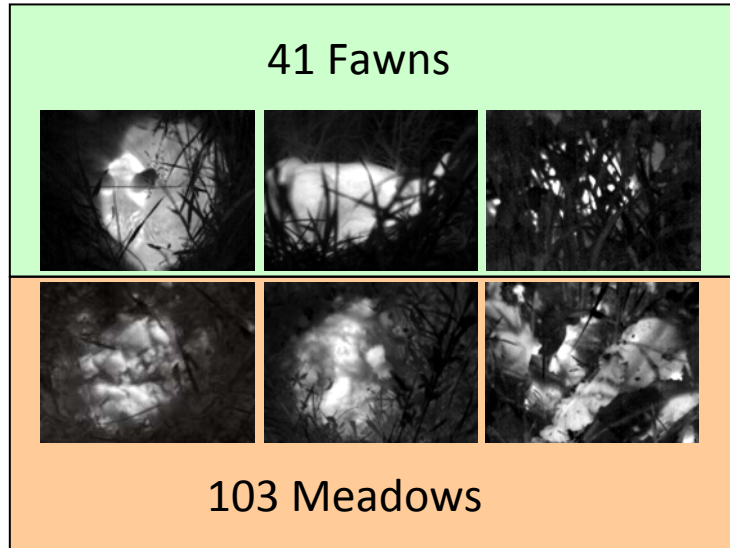


Classification accuracy: comparison with 6 other compression-based methods



Running times comparison for the top 3 methods

Example of NCD's Failure: Wild Animals Detection



The 3 missed detections (FCD)



Compressor used with NCD: LZW
Image size: 160x120

Confusion Matrices

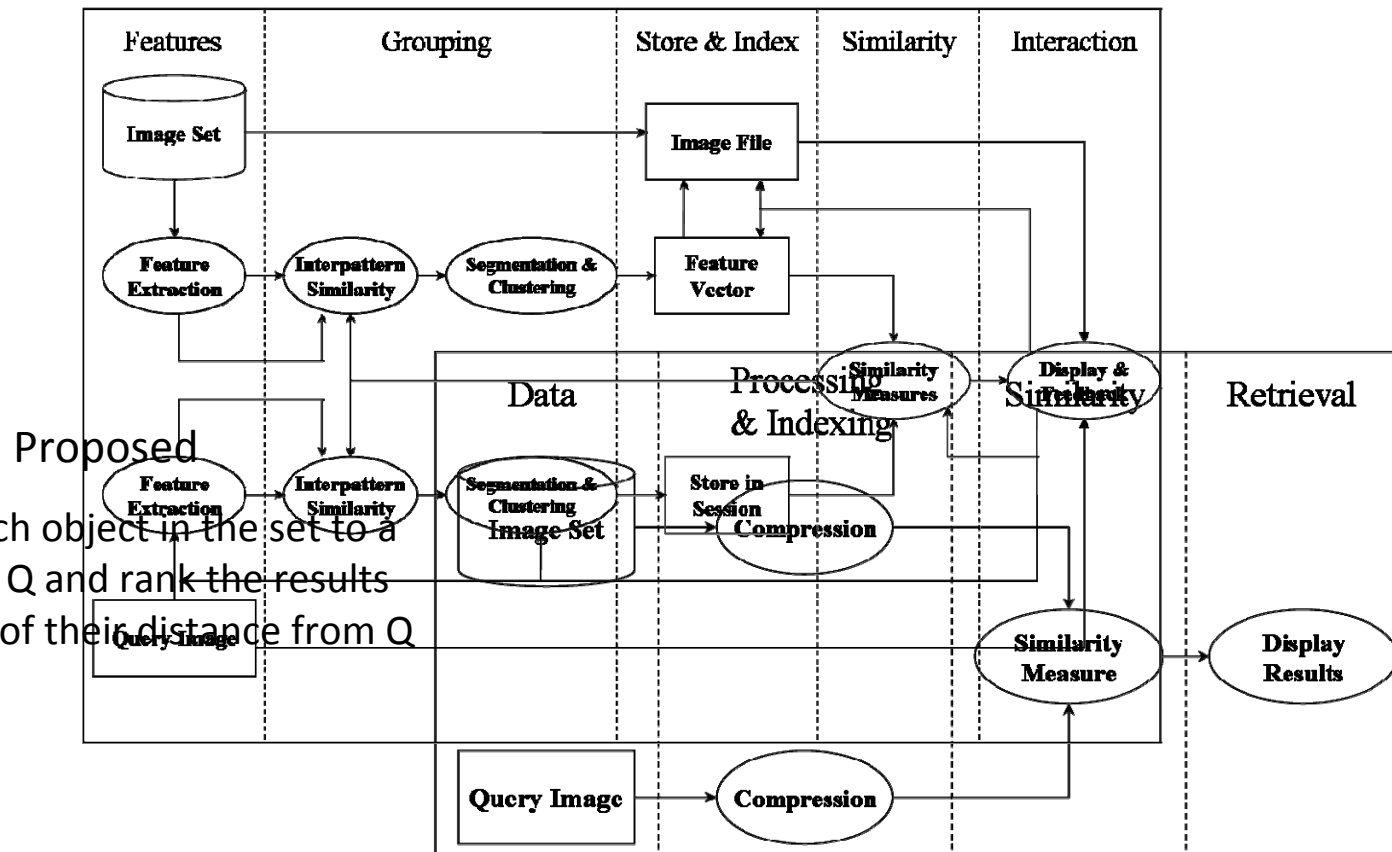
| | | Fawn | Meadow | Accuracy | Time |
|-----|--------|------|--------|----------|--------|
| FCD | Fawn | 38 | 3 | 97.9% | 58 sec |
| | Meadow | 0 | 103 | | |
| NCD | Fawn | 29 | 12 | 77.8% | 14 min |
| | Meadow | 20 | 83 | | |

Limited buffer size in the compressor and total loss of vertical texture causes NCD's performance to decrease!

Content-based image retrieval system

Classical (Smeulders, 2000)

Many steps and parameters to set



Proposed
Compare each object in the set to a query image Q and rank the results on the basis of their distance from Q

Applications: COREL Dataset



Competence Centre on Information Extraction
and Image Understanding for Earth Observation

Africans



Beach



Architecture



Buses



Dinosaurs



Elephants



Flowers



Horses



Mountains



Food



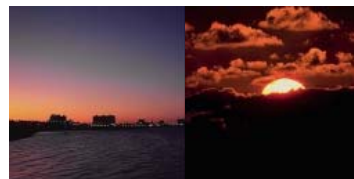
Caves



Postcards



Sunsets



Tigers



Women



1500 images, 15 classes, 100 images per class

Precision (P) vs. Recall (R) Evaluation

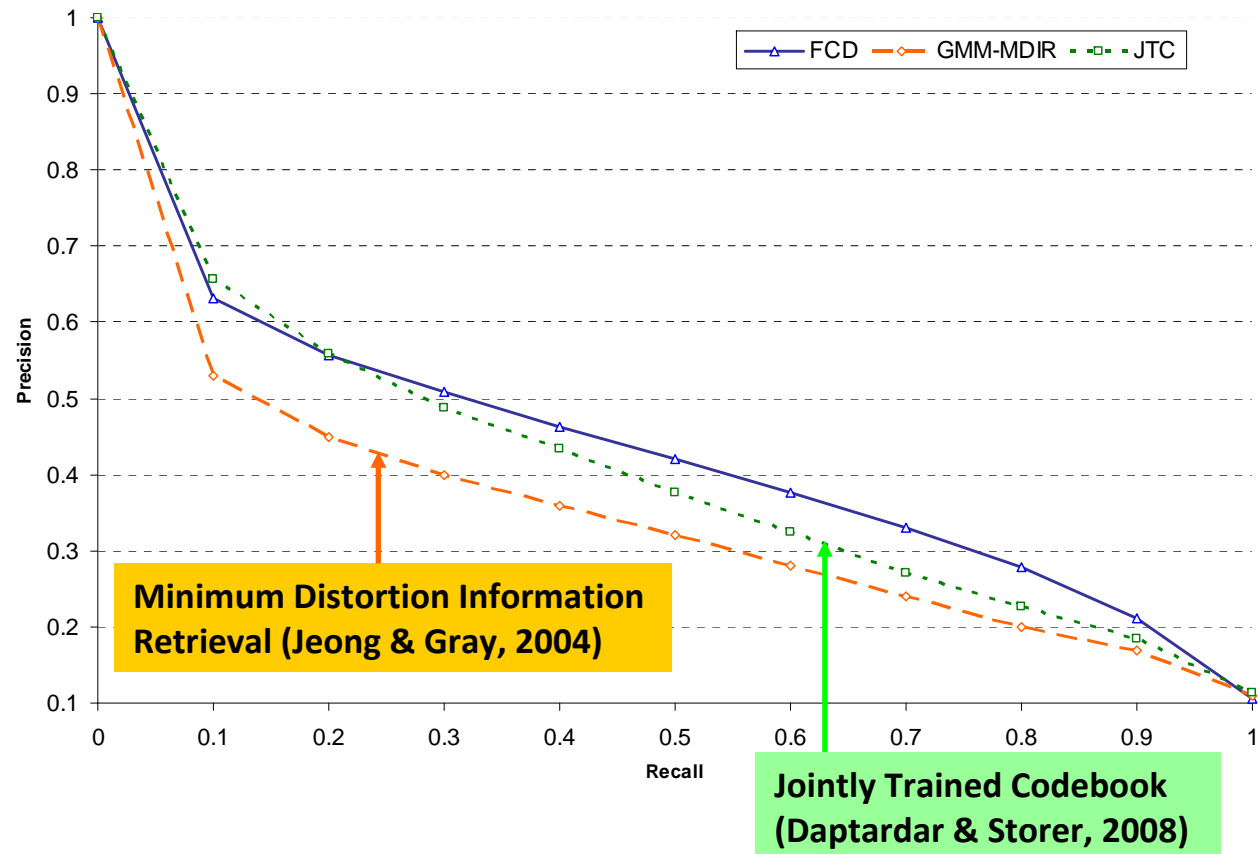
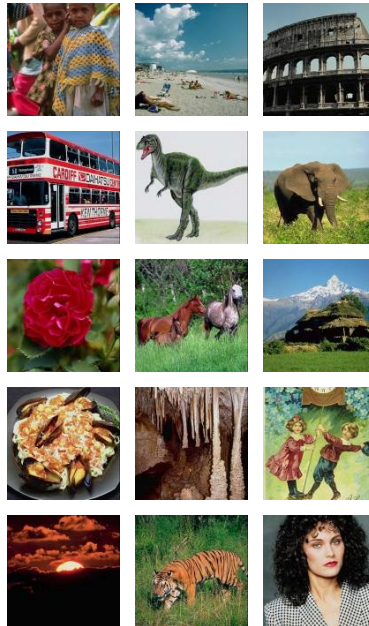
$$P = \frac{TP}{TP + FP}$$

True Positives
False Positives

$$R = \frac{TP}{TP + FN}$$

False Negatives

Competence Centre on Information Extraction and Image Understanding for Earth Observation



Running time: 18 min (images resampled to 64x64)



2 Processors (2GHz) + 2 GB RAM

Confusion Matrix

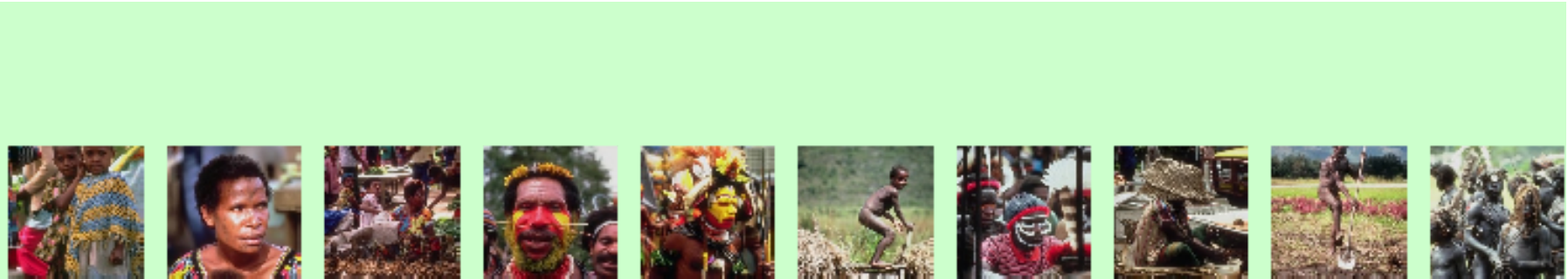
- Classification according to the minimum average distance from a class



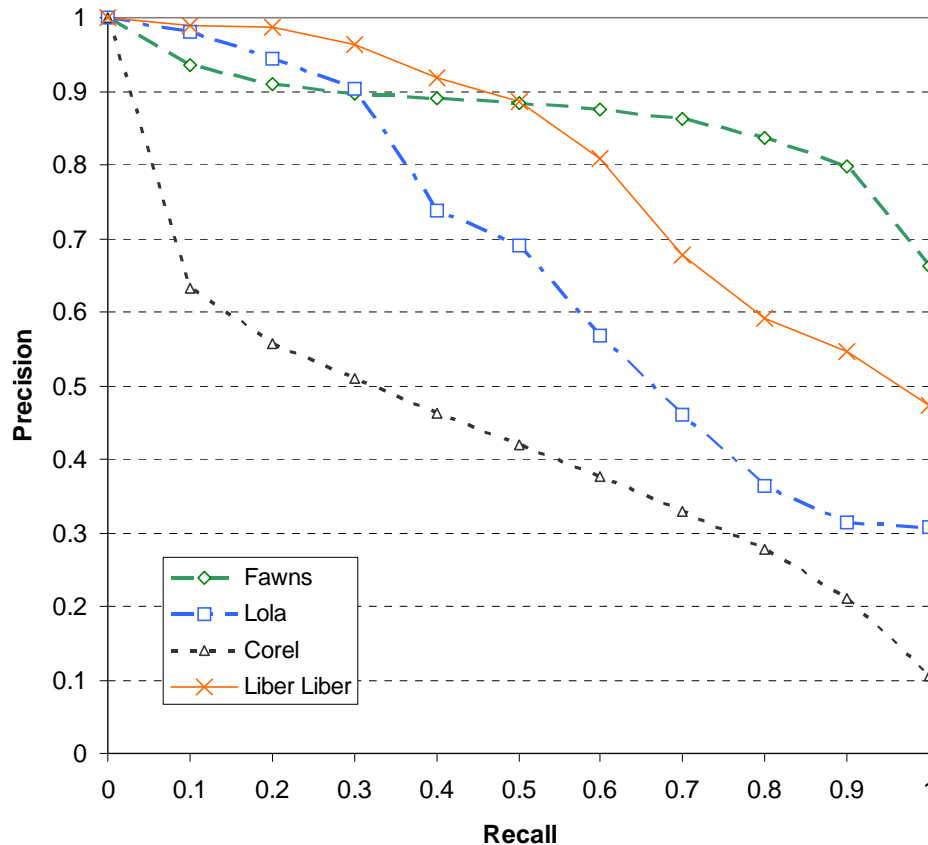
| | Afr. | Beach | Archit. | Bus. | Dinos. | Eleph. | Flow. | Hors. | Mount. | Food | Caves | Post. | Suns. | Tig. | Wom. |
|--------------|-----------|-----------|-----------|-----------|------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Africans | 90 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 8 | 0 |
| Beach | 12 | 43 | 8 | 14 | 0 | 1 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 18 | 0 |
| Architecture | 7 | 0 | 72 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 16 | 0 |
| Buses | 6 | 0 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Dinosaurs | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Elephants | 16 | 0 | 2 | 2 | 0 | 46 | 0 | 4 | 0 | 3 | 0 | 1 | 0 | 26 | 0 |
| Flowers | 6 | 0 | 3 | 1 | 0 | 0 | 83 | 1 | 0 | 3 | 0 | 0 | 0 | 3 | 0 |
| Horses | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| Mountains | 7 | 1 | 11 | 23 | 0 | 2 | 0 | 0 | 39 | 0 | 0 | 0 | 0 | 17 | 0 |
| Food | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 92 | 0 | 0 | 0 | 1 | 0 |
| Caves | 17 | 0 | 9 | 1 | 0 | 1 | 0 | 0 | 0 | 5 | 60 | 0 | 0 | 7 | 0 |
| Postcards | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 98 | 0 | 0 | 0 |
| Sunsets | 18 | 0 | 1 | 6 | 0 | 0 | 2 | 0 | 0 | 16 | 3 | 1 | 39 | 14 | 0 |
| Tigers | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 93 | 0 |
| Women | 35 | 0 | 0 | 6 | 2 | 0 | 0 | 0 | 0 | 20 | 4 | 0 | 0 | 5 | 28 |
| Avg Accuracy | 71.3% | | | | | | | | | | | | | | |

False alarms (?)

Typical images belonging to the class “Africans”



Estimation of the “complexity” of a dataset



Complexity

- Reduced
- Average
- High

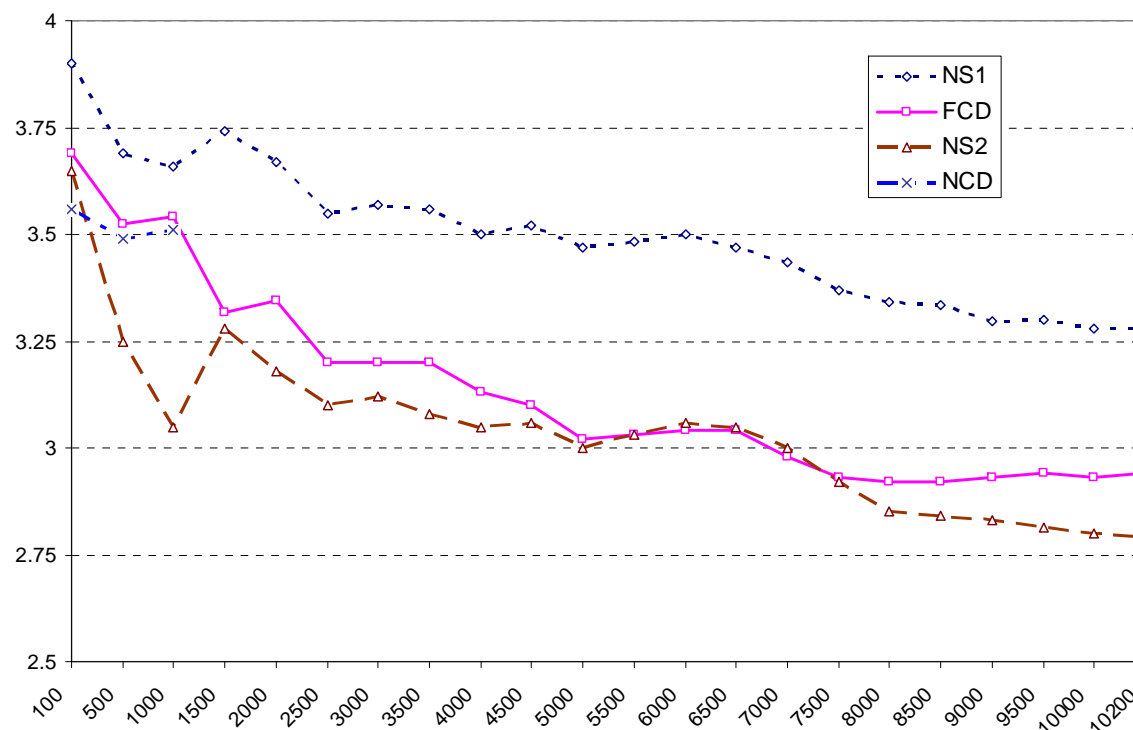
Rank

| Dataset | MAP score |
|-----------------|-----------|
| Fawns & Meadows | 0.872 |
| Liber Liber | 0.81 |
| Lola | 0.661 |
| Corel | 0.433 |

| Dataset | Classes | Objects | Content Diversity |
|-----------------|---------|---------|-------------------|
| Fawns & Meadows | 2 | 144 | Average |
| Lola | 19 | 164 | Average |
| Corel | 15 | 1500 | High |
| Liber Liber | 11 | 90 | Low |



A larger dataset and a comparison with state-of-the-art methods: Nister-Stewenius

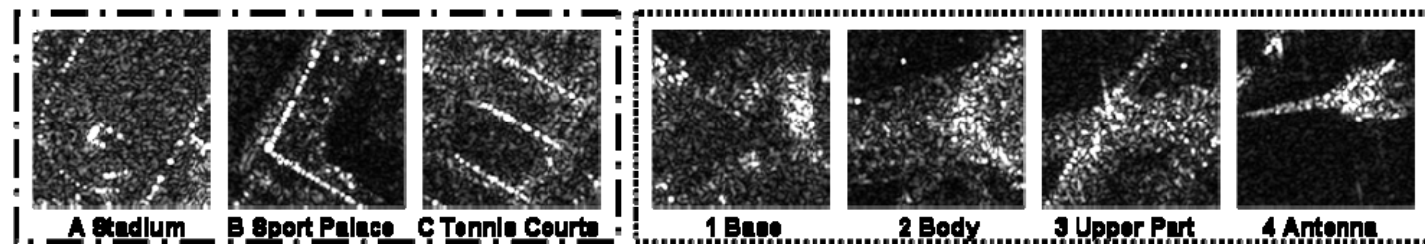
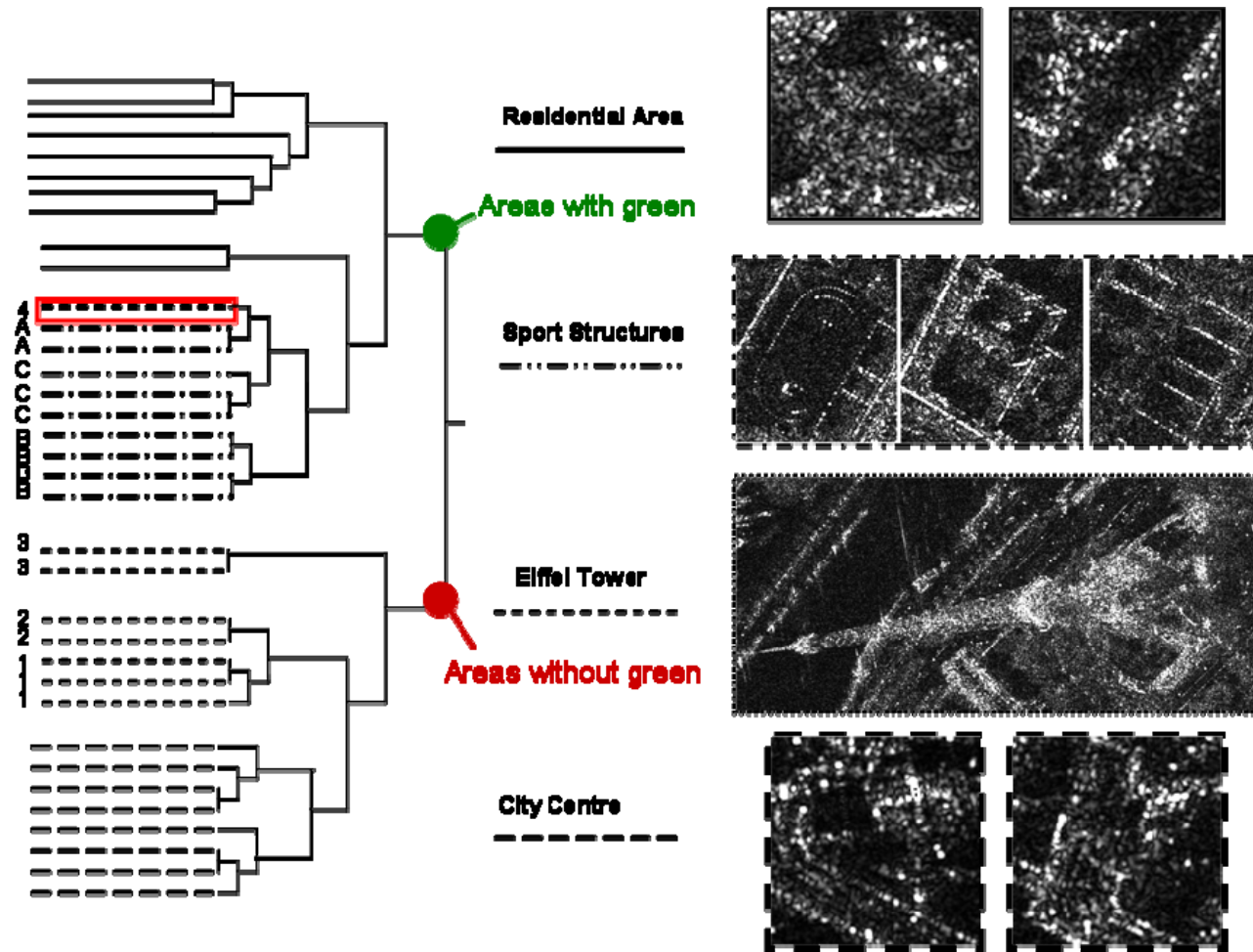


- 10200 images
- 2550 objects photographed under 4 points of view
- Score (from 1 to 4) represents the number of meaningful objects retrieved in the top-4
- SIFT-based NS1 and NS2 use different training sets and parameters settings, and yield different results
- FCD is independent from parameters
- Only 1000 images processed for NCD
- Query Time: 8 seconds

SAR Scene Hierarchical Clustering

32 TerraSAR-X subsets acquired over Paris

False Alarm



Outline

- The core: Compression-based similarity measures (CBSM)
- Theorectical Foundations
- Contributions: Theory
- Contributions: Applications and Experiments
- **Conclusions and Perspectives**

Summary



- Study and expansion of the Shannon-Kolmogorov parallel
- Bringing independent concepts into the frame
 - Compression-based “Relative entropy”
 - PRDC
- Fast Compression Distance based on explicit dictionaries
- Content-based Image Retrieval system
 - Parameter-free approach
 - Tests carried out on datasets 100 times larger than the ones used in the main works on the topic
- Estimation of the intrinsic complexity of an annotated dataset

Conclusions and Perspectives

- Compression-based similarity measures are not a magic wand!
 - Results obtained so far on small datasets could be misleading
 - On the larger datasets analyzed, results are often inferior to the state of the art
 - Open question: could they be somehow improved?

- Anyway their use in practical applications is justified
 - Overall satisfactory performance
 - Universally applicable
 - Simplicity in the implementation
 - Neither setting of parameters or any supervision from an expert required

- Future Perspectives
 - Integrate FCD in a DBMS
 - **May this help in applying CBSM to large datasets?**
 - **Would it be possible on its basis to define a semantic search engine?**
 - Analyze behaviour and advantages of lossy compression for CBIR systems



Thanks for your attention! 😊