Graph Mining and Graph Classification: Application to cadastral map analysis.

Romain Raveaux

L3I lab (EA 2118) – Université de La Rochelle

November 25-11, 2010
PhD Defense

Supervisors: J-M. Ogier
J-C. Burie

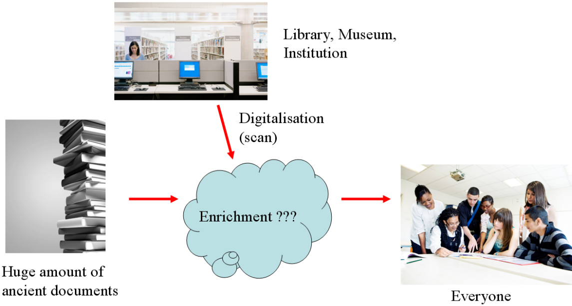
Outline

Ancient documents

1. Massive production of heterogeneous documents.
2. Societal issues and challenges
   - Heritage preservation
   - An open access to patrimonial knowledge
   - Historical enrichment
3. Digital library

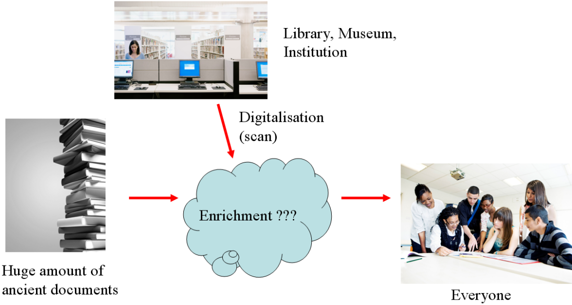
Ancient documents

1. Massive production of heterogeneous documents.
2. Societal issues and challenges
   - Heritage preservation
   - An open access to patrimonial knowledge
   - Historical enrichment
3. Digital library



Ancient documents

1. Massive production of heterogeneous documents.
2. Societal issues and challenges
   - Heritage preservation
   - An open access to patrimonial knowledge
   - Historical enrichment
3. Digital library



Digital library

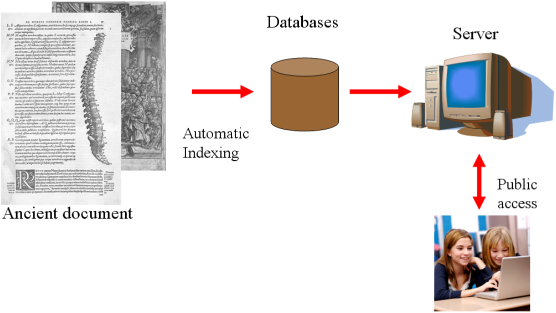
Textual documents

- Manual insertion of meta-data
- Automatic indexing
  - OCR for old characters (DEBORA project [Boucher00])
  - Structure indexing (AGORA [Ramel06])
  - Texture indexing [Journet 08]

Digital library

Textual documents

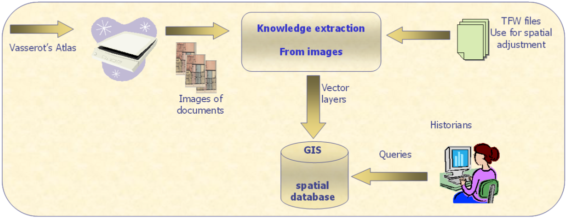
- Manual insertion of meta-data
- Automatic indexing
  - OCR for old characters (DEBORA project [Boucher00])
  - Structure indexing (AGORA [Ramel06])
  - Texture indexing [Journet 08]



Digital library

Graphic documents:

- Automatic indexing
  - Symbol descriptor [T-O Nguyen 08]
  - Relational indexing in line-drawing images [Rusiñol 10]
  - Drop cap indexing [Uttama 05] [Coustaty 09]
  - Map indexing (ALPAGE project)



ALPAGE project

1. ALPAGE (diachronic analysis of the Paris urban area: a geomatic approach)
2. Supported by the ANR (National Research Agency)
3. An association of 4 laboratories.
   - **LAMOP** of Paris-1, carrying the project, which includes historians, specialists in urban history and digital tools.
   - **LIENSS** of La Rochelle: geographers specialized in geomatics.
   - **ArScAn** in Nanterre bringing together archaeologists and geomaticians skilled in GIS and archeology of the parisian area.
   - **L3i** of La Rochelle,comprised of IT scientists specialized in pattern recognition and vectorization.
4. Objective: To build a geographic information system (GIS) about the pre-industrial Parisian area.

ALPAGE project

1. ALPAGE (diachronic analysis of the Paris urban area: a geomatic approach)
2. Supported by the ANR (National Research Agency)
3. An association of 4 laboratories.
   - **LAMOP** of Paris-1, carrying the project, which includes historians, specialists in urban history and digital tools.
   - **LIENSS** of La Rochelle: geographers specialized in geomatics.
   - **ArScAn** in Nanterre bringing together archaeologists and geomaticians skilled in GIS and archeology of the parisian area.
   - **L3i** of La Rochelle,comprised of IT scientists specialized in pattern recognition and vectorization.
4. Objective: To build a geographic information system (GIS) about the pre-industrial Parisian area.

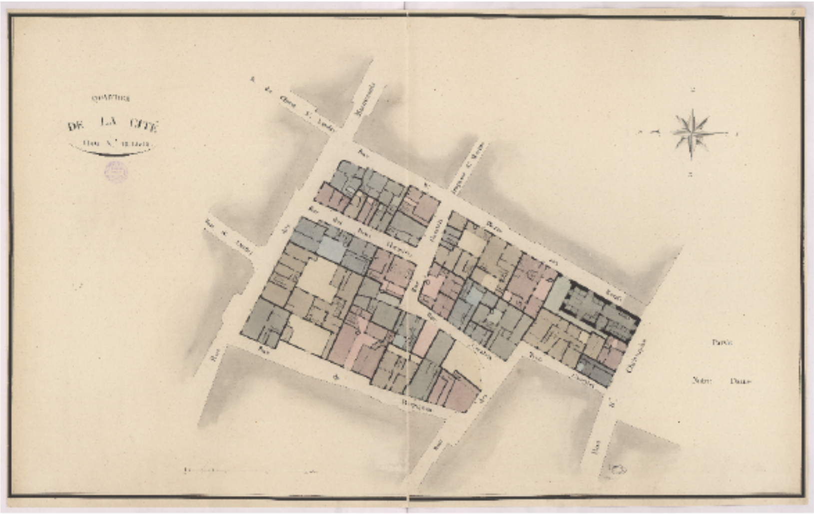
ALPAGE project

1. ALPAGE (diachronic analysis of the Paris urban area: a geomatic approach)
2. Supported by the ANR (National Research Agency)
3. An association of 4 laboratories.
   - **LAMOP** of Paris-1, carrying the project, which includes historians, specialists in urban history and digital tools.
   - **LIENSS** of La Rochelle: geographers specialized in geomatics.
   - **ArScAn** in Nanterre bringing together archaeologists and geomaticians skilled in GIS and archeology of the parisian area.
   - **L3i** of La Rochelle,comprised of IT scientists specialized in pattern recognition and vectorization.
4. Objective: To build a geographic information system (GIS) about the pre-industrial Parisian area.

ALPAGE project

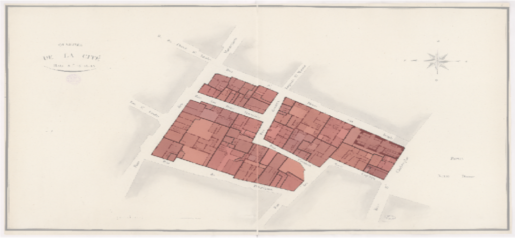
1. ALPAGE (diachronic analysis of the Paris urban area: a geomatic approach)
2. Supported by the ANR (National Research Agency)
3. An association of 4 laboratories.
   - **LAMOP** of Paris-1, carrying the project, which includes historians, specialists in urban history and digital tools.
   - **LIENSS** of La Rochelle: geographers specialized in geomatics.
   - **ArScAn** in Nanterre bringing together archaeologists and geomaticians skilled in GIS and archeology of the parisian area.
   - **L3i** of La Rochelle,comprised of IT scientists specialized in pattern recognition and vectorization.
4. Objective: To build a geographic information system (GIS) about the pre-industrial Parisian area.

ALPAGE: Raster to Polygon



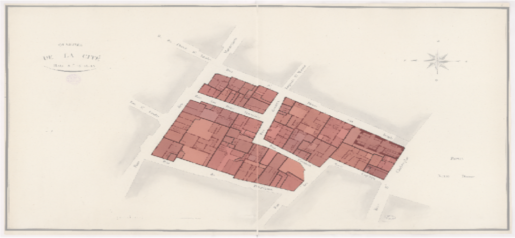
ALPAGE: Raster to Polygon

- Information retrieval
  - Quarter
  - Parcel
- Parcel Polygonization

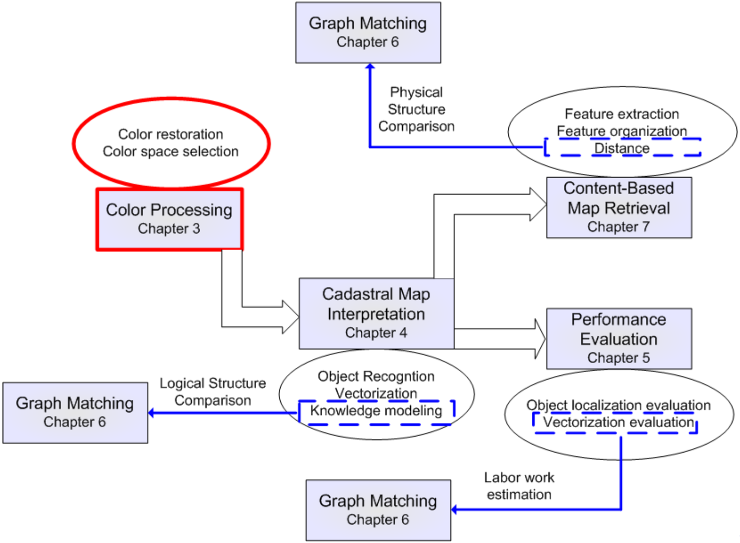


ALPAGE: Raster to Polygon

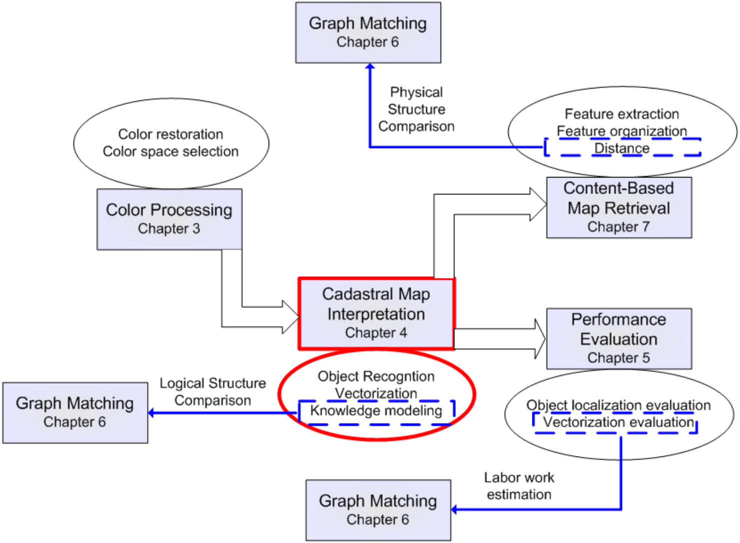
- Information retrieval
  - Quarter
  - Parcel
- Parcel Polygonization



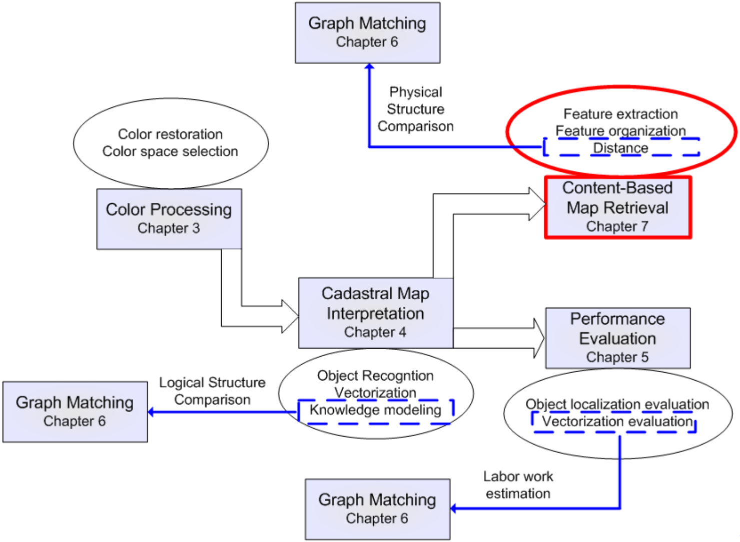
Overall methodology of our system



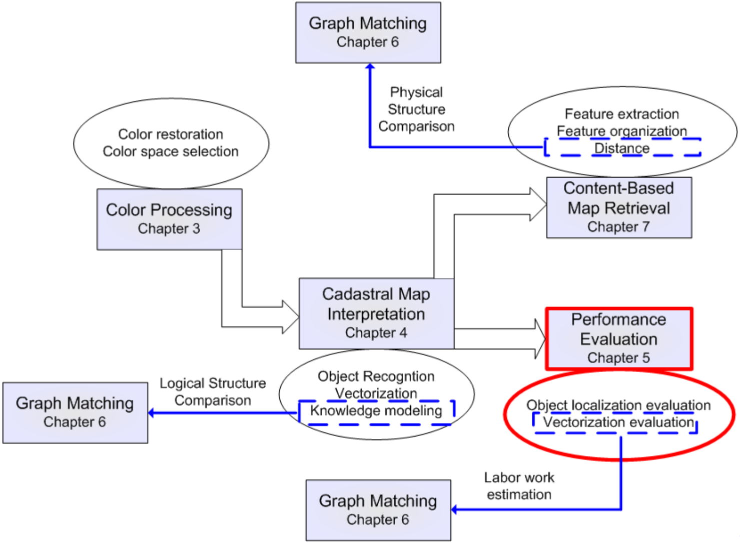
Overall methodology of our system



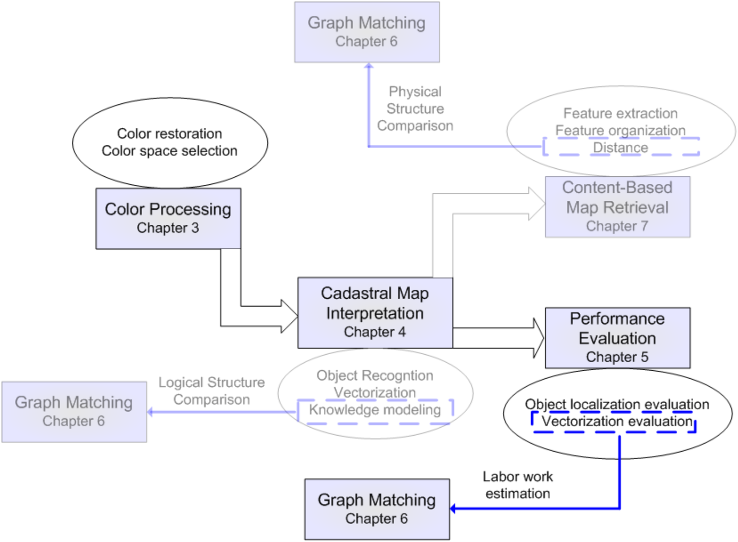
Overall methodology of our system



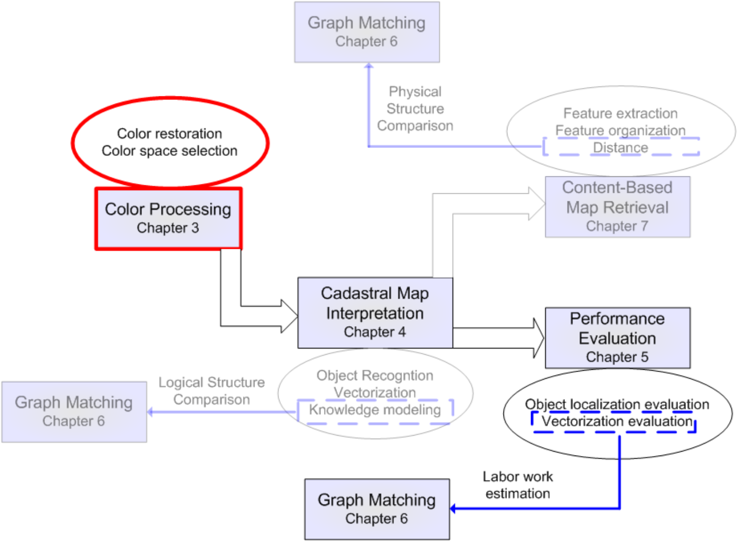
Overall methodology of our system



Overall methodology of our system



Overall methodology of our system

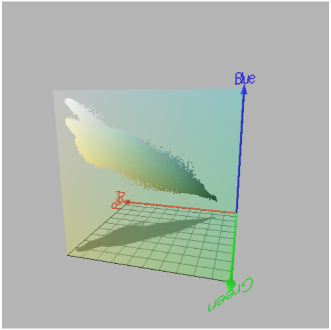


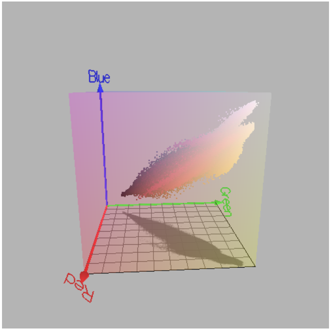
Input images

- Time due degradation
- Under-saturated images
  - More washed out, as in pastels
- Color restoration
  - Non-uniform increasing of the saturation

Input images

- Time due degradation
- Under-saturated images
  - More washed out, as in pastels
- Color restoration
  - Non-uniform increasing of the saturation





Input images

- Time due degradation
- Under-saturated images
    - More washed out, as in pastels
- Color restoration
    - Non-uniform increasing of the saturation

Color enhancement based on PCA

- Independent system axis: $Y = V(X - \mu) \quad X = \begin{vmatrix} R \\ G \\ B \end{vmatrix}$
  - V are the eigenvectors of the covariance matrix.
  - μ is the mean vector.
- Data extension in the direction of the main factorial axis.

$$Y' = KY$$

$$K = \begin{vmatrix} k1 & 0 & 0 \\ 0 & k2 & 0 \\ 0 & 0 & k3 \end{vmatrix}$$

Color enhancement based on PCA

- Independent system axis: $Y = V(X - \mu) \quad X = \begin{vmatrix} R \\ G \\ B \end{vmatrix}$
  - V are the eigenvectors of the covariance matrix.
  - μ is the mean vector.
- Data extension in the direction of the main factorial axis.

$$Y' = KY$$

$$K = \begin{vmatrix} k1 & 0 & 0 \\ 0 & k2 & 0 \\ 0 & 0 & k3 \end{vmatrix}$$

Conventional representation

- Difference color spaces:
  - Primary based system: *RGB*
  - Perceptual color space: $L^*a^*b^*$
  - Luminance – Chrominance representation: $AC1C2$
  - Independ axis system: $I1I2I3$
- A set of color components:

$$C = \{C_i\}_{i=1}^{N} = \{R, G, B, I1, I2, I3, L^*, a^*, b^*, ...\}$$

with Card(C)=25

The choice of a color space turns into a feature selection problem

Conventional representation

- Difference color spaces:
  - Primary based system: RGB
  - Perceptual color space: $L^* a^* b^*$
  - Luminance – Chrominance representation: $AC1C2$
  - Independ axis system: $I1I2I3$
- A set of color components:

$$C = \{C_i\}_{i=1}^{N} = \{R, G, B, I1, I2, I3, L^*, a^*, b^*, \ldots\}$$

with Card(C)=25

The choice of a color space turns into a feature selection problem

Conventional representation

- Difference color spaces:
  - Primary based system: RGB
  - Perceptual color space: $L^* a^* b^*$
  - Luminance – Chrominance representation: $AC1C2$
  - Independ axis system: $I1I2I3$
- A set of color components:

$$C = \{C_i\}_{i=1}^{N} = \{R, G, B, I1, I2, I3, L^*, a^*, b^*, ...\}$$

with Card(C)=25

The choice of a color space turns into a feature selection problem

Conventional representation

- Difference color spaces:
  - Primary based system: RGB
  - Perceptual color space: $L^*a^*b^*$
  - Luminance – Chrominance representation: $AC1C2$
  - Independ axis system: $I1I2I3$
- A set of color components:

$$C = \{C_i\}_{i=1}^{N} = \{R, G, B, I1, I2, I3, L^*, a^*, b^*, \ldots\}$$

with Card(C)=25

The choice of a color space turns into a feature selection problem

Conventional representation

- Difference color spaces:
  - Primary based system: RGB
  - Perceptual color space: $L^* a^* b^*$
  - Luminance – Chrominance representation: $AC1C2$
  - Independ axis system: $I1I2I3$
- A set of color components:

$$C = \{C_i\}_{i=1}^{N} = \{R, G, B, I1, I2, I3, L^*, a^*, b^*, ...\}$$

with Card(C)=25

The choice of a color space turns into a feature selection problem

Conventional representation

- Difference color spaces:
  - Primary based system:RGB
  - Perceptual color space: $L^* a^* b^*$
  - Luminance – Chrominance representation: $AC1C2$
  - Independ axis system: $I1I2I3$
- A set of color components:

$$C = \{C_i\}_{i=1}^{N} = \{R, G, B, I1, I2, I3, L^*, a^*, b^*, ...\}$$

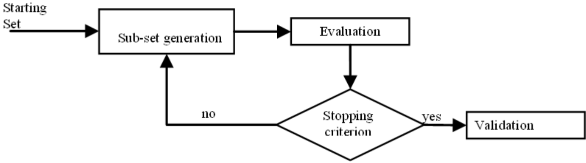
with Card(C)=25

The choice of a color space turns into a feature selection problem

Feature selection

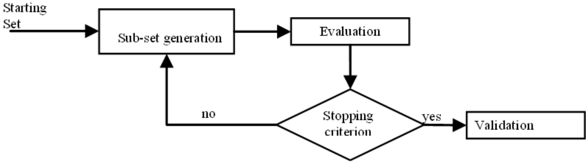
- Find $K \subset C$ with $Card(K) = 3$
- Criteria: Maximization of a classification rate
- Classification: 1-NN
- Search algorithm:

Name	Type	Searching algorithm
CFS	Filter	Greedy stepwise
DHCS	Filter	Ranker
GACS	Wrapper	Genetic Algorithm
OneRS	Wrapper	Ranker



Feature selection

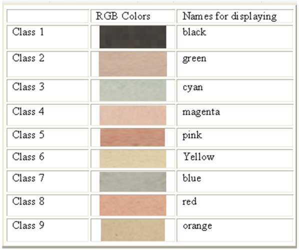
- Find $K \subset C$ with $Card(K) = 3$
- Criteria: Maximization of a classification rate
- Classification: 1-NN
- Search algorithm:



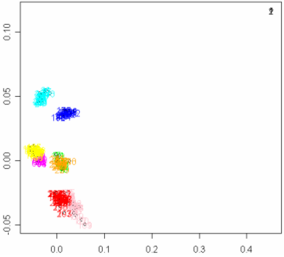
Name	Type	Searching algorithm
CFS	Filter	Greedy stepwise
DHCS	Filter	Ranker
GACS	Wrapper	Genetic Algorithm
OneRS	Wrapper	Ranker

Feature selection

- Find $K \subset C$ with $Card(K) = 3$
- Criteria: Maximization of a classification rate
- Classification: 1-NN
- Search algorithm:



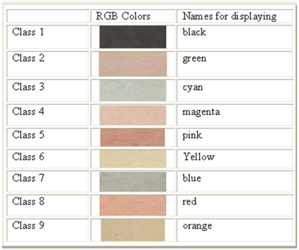
	RGB Colors	Names for displaying
Class 1		black
Class 2		green
Class 3		cyan
Class 4		magenta
Class 5		pink
Class 6		Yellow
Class 7		blue
Class 8		red
Class 9		orange



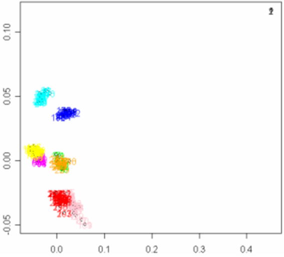
Name	Type	Searching algorithm
CFS	Filter	Greedy stepwise
DHCS	Filter	Ranker
GACS	Wrapper	Genetic Algorithm
OneRS	Wrapper	Ranker

Feature selection

- Find $K \subset C$ with $Card(K) = 3$
- Criteria: Maximization of a classification rate
- Classification: 1-NN
- Search algorithm:



	RGB Colors	Names for displaying
Class 1		black
Class 2		green
Class 3		cyan
Class 4		magenta
Class 5		pink
Class 6		Yellow
Class 7		blue
Class 8		red
Class 9		orange



Name	Type	Searching algorithm
CFS	Filter	Greedy stepwise
DHCS	Filter	Ranker
GACS	Wrapper	Genetic Algorithm
OneRS	Wrapper	Ranker

Vectorial gradient

- Edge detection:
  - Di Zenzo's method
- Vectorial gradient in K

$$\begin{cases} a = & (G_x^{K1})^2 + (G_x^{K2})^2 + (G_x^{K3})^2 \\ b = & G_x^{K1} G_y^{K1} + G_x^{K2} G_y^{K2} + G_x^{K3} G_y^{K3} \\ c = & (G_y^{K1})^2 + (G_y^{K2})^2 + (G_y^{K3})^2 \end{cases}$$

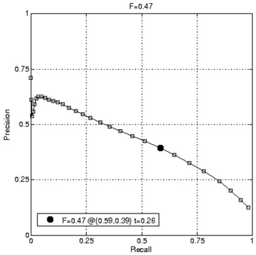
- Segmentation results
  - on Berkeley benchmark
  - Slight improvement

Vectorial gradient

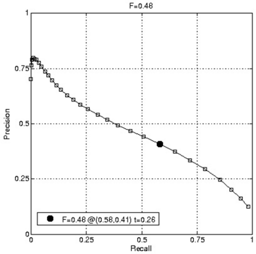
- Edge detection:
  - Di Zenzo's method
- Vectorial gradient in K

$$\begin{cases} a = & (G_x^{K1})^2 + (G_x^{K2})^2 + (G_x^{K3})^2 \\ b = & G_x^{K1} G_y^{K1} + G_x^{K2} G_y^{K2} + G_x^{K3} G_y^{K3} \\ c = & (G_y^{K1})^2 + (G_y^{K2})^2 + (G_y^{K3})^2 \end{cases}$$

- Segmentation results
  - on Berkeley benchmark
  - Slight improvement

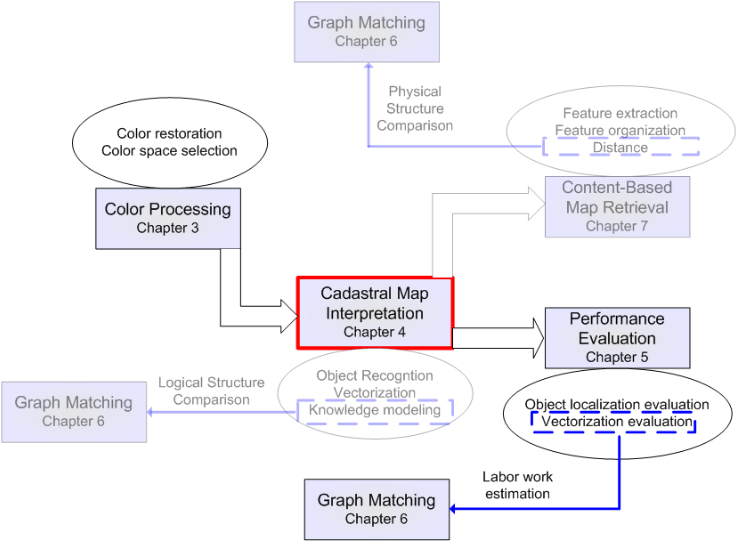


RGB



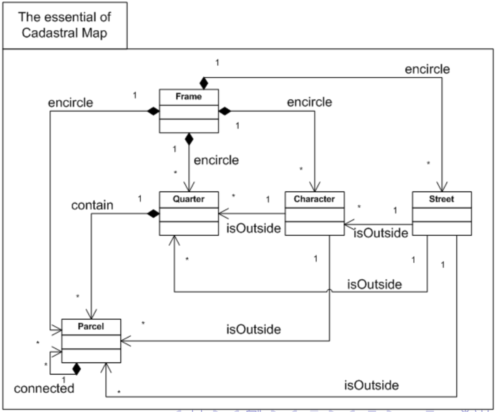
K

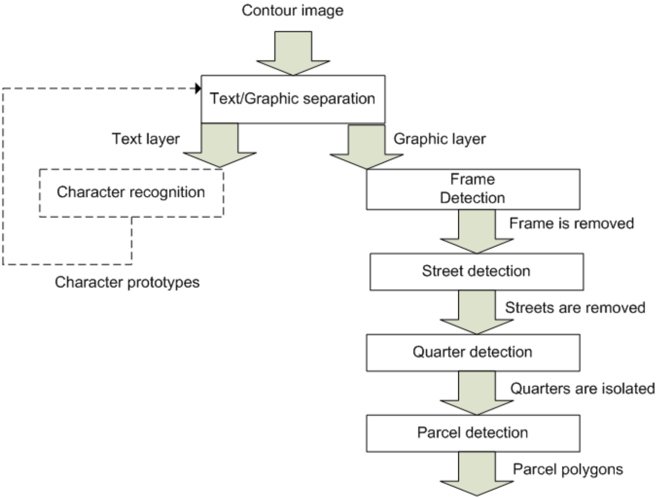
Main steps


Graph Matching
Chapter 6
Physical
Structure
Comparison
Color restoration
Color space selection
Feature extraction
Feature organization
Distance
Content-Based
Map Retrieval
Chapter 7
Color Processing
Chapter 3
Cadastral Map
Interpretation
Chapter 4
Performance
Evaluation
Chapter 5
Logical Structure
Comparison
Object Recogntion
Vectorization
Knowledge modeling
Graph Matching
Chapter 6
Object localization evaluation
Vectorization evaluation
Labor work
estimation
Graph Matching
Chapter 6

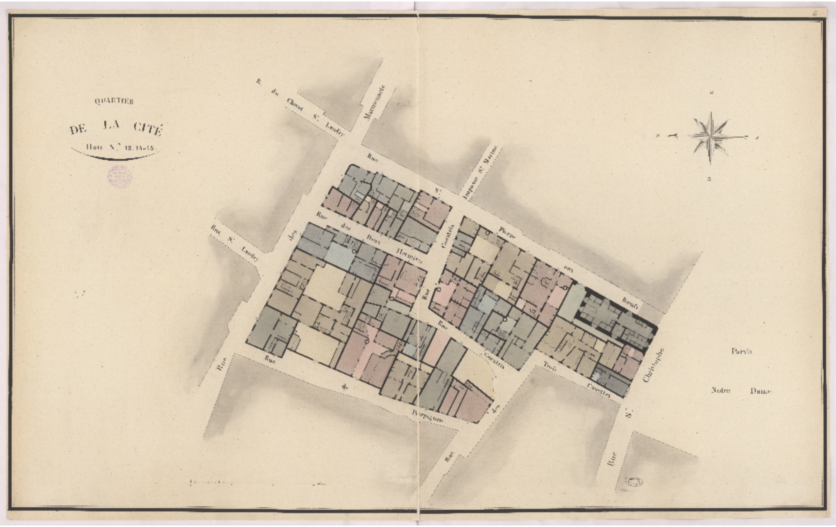
Modeling

- Logical structure
  - To identify the map elements

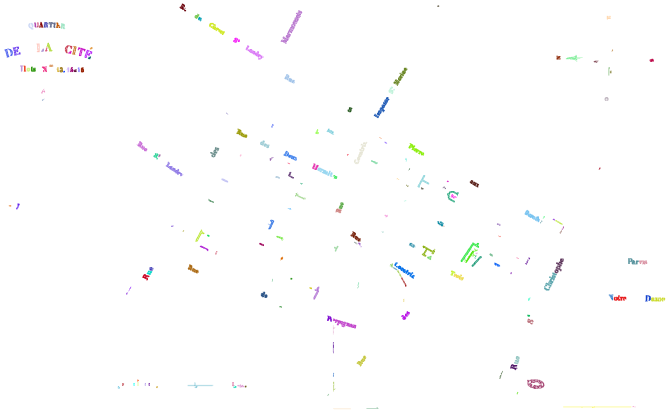



Contour image
Text/Graphic separation
Text layer
Graphic layer
Character recognition
Character prototypes
Frame
Detection
Frame is removed
Street detection
Streets are removed
Quarter detection
Quarters are isolated
Parcel detection
Parcel polygons

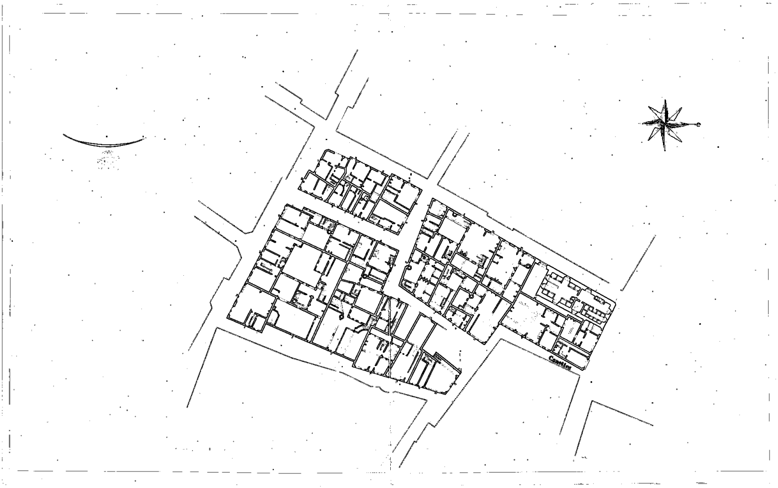
Original image



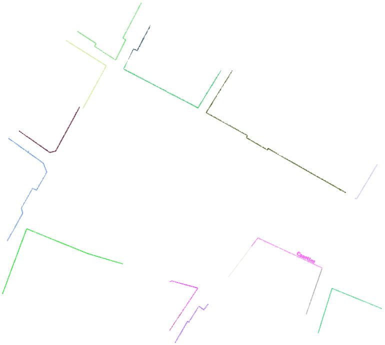
Text layer



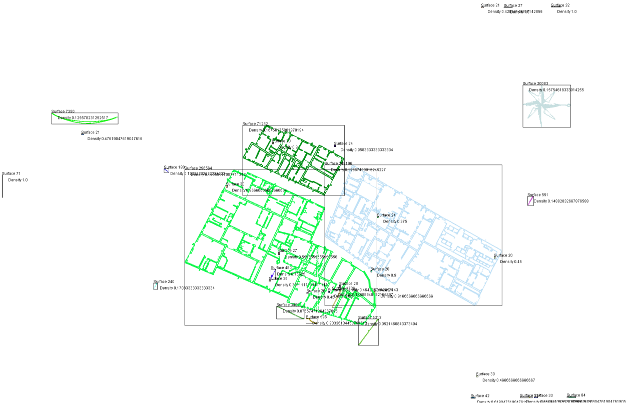
Graphic layer



Street layer

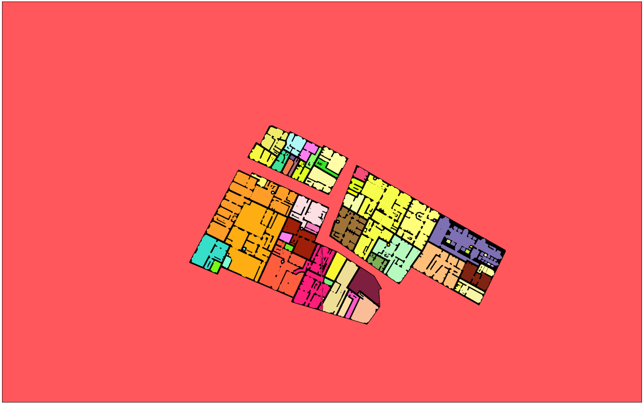


Graphic minus Street



Quarters

White connected components



Black layer remover: Median axis

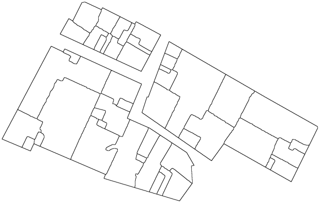
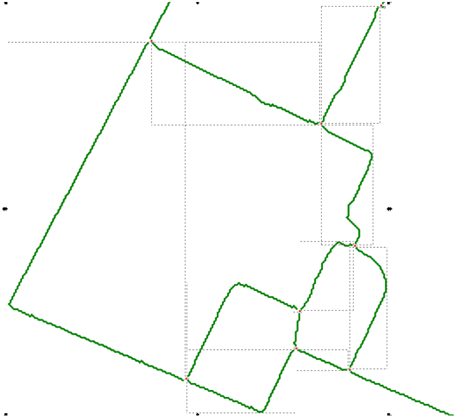
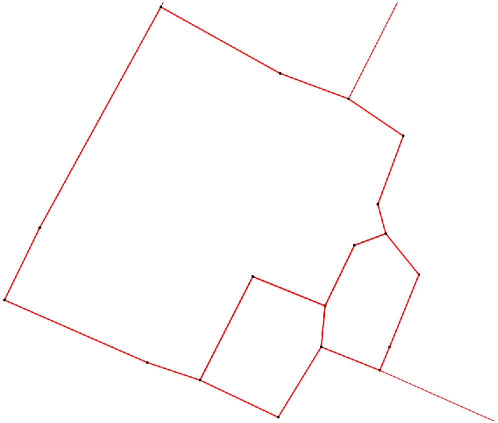


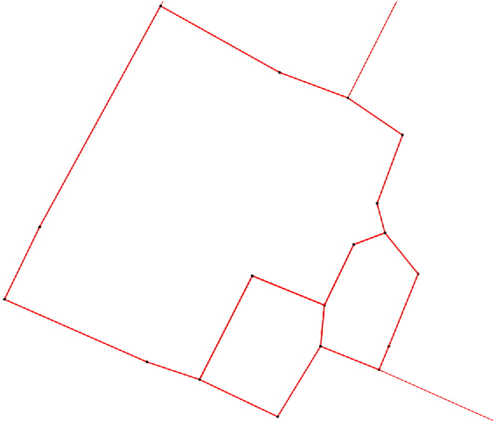
Image chaining



Polygonal approximation

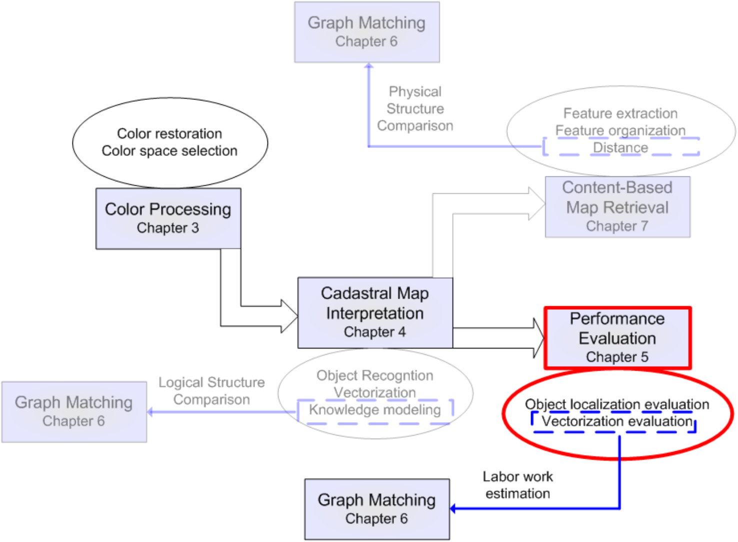


Raster to polygon system

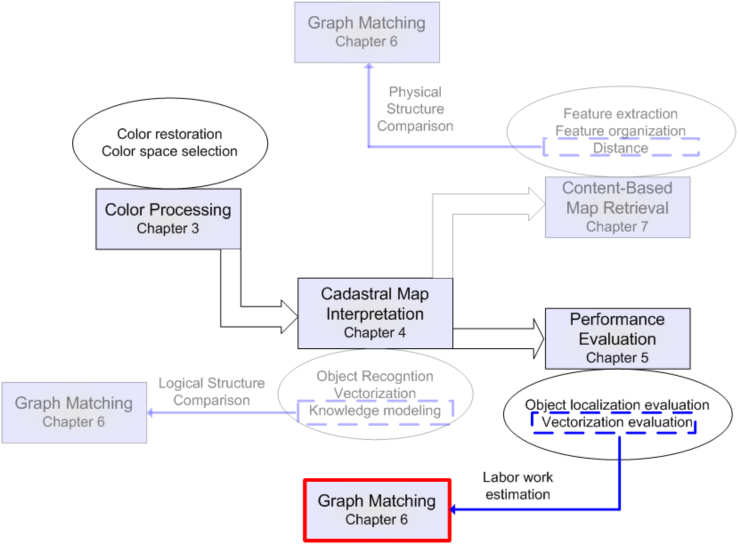


- Polygon production
- We need to evaluate it

Main steps


Graph Matching
Chapter 6
Physical
Structure
Comparison
Color restoration
Color space selection
Feature extraction
Feature organization
Distance
Color Processing
Chapter 3
Content-Based
Map Retrieval
Chapter 7
Cadastral Map
Interpretation
Chapter 4
Performance
Evaluation
Chapter 5
Graph Matching
Chapter 6
Logical Structure
Comparison
Object Recogntion
Vectorization
Knowledge modeling
Object localization evaluation
Vectorization evaluation
Labor work
estimation
Graph Matching
Chapter 6

Main steps


Graph Matching
Chapter 6
Physical
Structure
Comparison
Feature extraction
Feature organization
Distance
Color restoration
Color space selection
Color Processing
Chapter 3
Content-Based
Map Retrieval
Chapter 7
Cadastral Map
Interpretation
Chapter 4
Performance
Evaluation
Chapter 5
Graph Matching
Chapter 6
Logical Structure
Comparison
Object Recogntion
Vectorization
Knowledge modeling
Object localization evaluation
Vectorization evaluation
Labor work
estimation
Graph Matching
Chapter 6

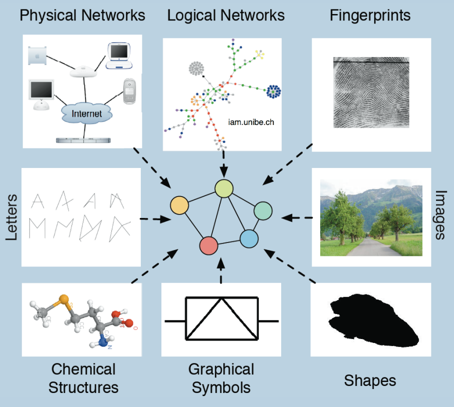
Graphs are everywhere...

Graphs in Reality

- Graphs model objects and their relationships.
- Also referred to as networks.
- All common data structures can be modeled as graphs.

How similar are two graphs?

- Graph similarity is the central problem for all learning tasks such as clustering and classification on graphs.



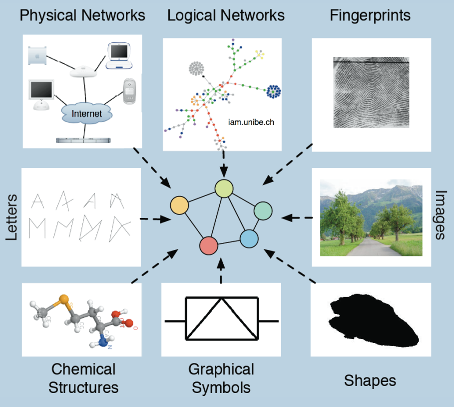
Graphs are everywhere...

Graphs in Reality

- Graphs model objects and their relationships.
- Also referred to as networks.
- All common data structures can be modeled as graphs.

How similar are two graphs?

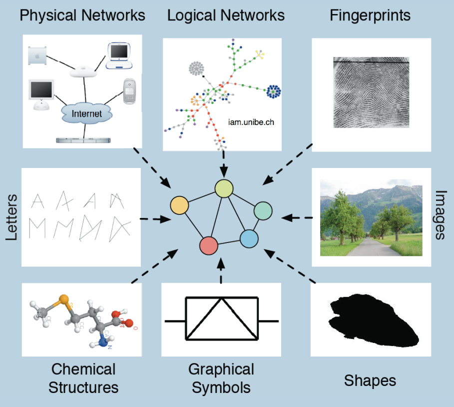
- Graph similarity is the central problem for all learning tasks such as clustering and classification on graphs.



Graphs are everywhere...

Graphs in Reality

- Graphs model objects and their relationships.
- Also referred to as networks.
- All common data structures can be modeled as graphs.



How similar are two graphs?

- Graph similarity is the central problem for all learning tasks such as clustering and classification on graphs.

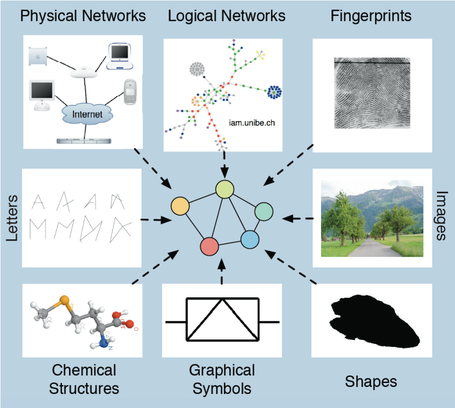
Graphs are everywhere...

Graphs in Reality

- Graphs model objects and their relationships.
- Also referred to as networks.
- All common data structures can be modeled as graphs.

How similar are two graphs?

- Graph similarity is the central problem for all learning tasks such as clustering and classification on graphs.



From the beginning...

Definition and notation of a graph:

Definition

Let L_V and L_E denote the set of node and edge labels, respectively. A labeled graph G is a 4-tuple $G = (V, E, \mu, \xi)$, where

- V is the set of nodes,
- $E \subseteq V \times V$ is the set of edges
- $\mu : V \rightarrow L_V$ is a function assigning labels to the nodes, and
- $\xi : E \rightarrow L_E$ is a function assigning labels to the edges.

- Let $G_1 = (V_1, E_1, \mu_1, \xi_1)$ be the source graph
- And $G_2 = (V_2, E_2, \mu_2, \xi_2)$ the target graph
- With $V_1 = (u_1, ..., u_n)$ and $V_2 = (v_1, ..., v_m)$ respectively

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \longrightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete

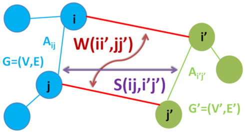
Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism $iff (f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete



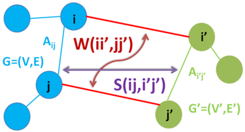
Graph isomorphism

Graph isomorphism

- Find a mapping $f : V_1 \rightarrow V_2$
- i.e. $x, y \in V_1 \Rightarrow (x, y) \in E_1$
- f is an isomorphism *iff* $(f(x), f(y))$ is an edge of G_2.
- No polynomial-time algorithm is known for graph isomorphism
- Neither it is known to be NP-complete

Subgraph isomorphism

- Means finding a subgraph G_3 of G_2 such that G_1 and G_3 are isomorphic.
- Subgraph isomorphism is NP-complete



Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Error-tolerant graph isomorphism

- Exact graph matching is useless in many computer vision applications
- Concerning graph matching under noise and distortion
- The matching incorporates an error model to identify the distortions which make one graph a distorted version of the other

Problems for real world applications

- Error-tolerant
- To measure the similarity of two graphs.
- Runtime may grow exponentially with number of nodes
- This is an enormous problem for large datasets of graphs

Wanted: Polynomial-time similarity measure for graphs

Problem statement

A dissimilarity measure is a function : $d : X \times X \rightarrow \mathbb{R}$
where X is the representation space for the object description.

1. non-negativity: $d(x, y) \geq 0$
2. uniqueness: $d(x, y) = 0 \Rightarrow x = y$
3. symmetry: $d(x, y) = d(y, x)$

Criteria for a good graph measure of similarity

- Expressive
- Efficient to compute
- Applicable to wide range of graphs

Problem statement

A dissimilarity measure is a function : $d : X \times X \rightarrow \mathbb{R}$
where X is the representation space for the object description.

1. non-negativity: $d(x, y) \geq 0$
2. uniqueness: $d(x, y) = 0 \Rightarrow x = y$
3. symmetry: $d(x, y) = d(y, x)$

Criteria for a good graph measure of similarity

- Expressive
- Efficient to compute
- Applicable to wide range of graphs

Problem statement

A dissimilarity measure is a function : $d : X \times X \rightarrow \mathbb{R}$ where X is the representation space for the object description.

1. non-negativity: $d(x, y) \geq 0$
2. uniqueness: $d(x, y) = 0 \Rightarrow x = y$
3. symmetry: $d(x, y) = d(y, x)$

Criteria for a good graph measure of similarity

- Expressive
- Efficient to compute
- Applicable to wide range of graphs

Problem statement

A dissimilarity measure is a function : $d : X \times X \rightarrow \mathbb{R}$
where X is the representation space for the object description.

1. non-negativity: $d(x, y) \geq 0$
2. uniqueness: $d(x, y) = 0 \Rightarrow x = y$
3. symmetry: $d(x, y) = d(y, x)$

Criteria for a good graph measure of similarity

- Expressive
- Efficient to compute
- Applicable to wide range of graphs

Problem statement

A dissimilarity measure is a function : $d : X \times X \rightarrow \mathbb{R}$
where X is the representation space for the object description.

1. non-negativity: $d(x, y) \geq 0$
2. uniqueness: $d(x, y) = 0 \Rightarrow x = y$
3. symmetry: $d(x, y) = d(y, x)$

Criteria for a good graph measure of similarity

- Expressive
- Efficient to compute
- Applicable to wide range of graphs

Problem statement

A dissimilarity measure is a function : $d : X \times X \rightarrow \mathbb{R}$
where X is the representation space for the object description.

1. non-negativity: $d(x, y) \geq 0$
2. uniqueness: $d(x, y) = 0 \Rightarrow x = y$
3. symmetry: $d(x, y) = d(y, x)$

Criteria for a good graph measure of similarity

- Expressive
- Efficient to compute
- Applicable to wide range of graphs

Comparison between Classical Graph-Matching Methods

	Graph Isomorphism	Subgraph Isomorphism	Error-tolerant Subgraph Isomorphism	Optimal	Complexity Class	Key References
Backtrack tree search	Yes	Yes	No	Yes	NP	
Forward checking	Yes	Yes	No	Yes	NP	[32]
Discrete relaxation	Yes	Yes	Yes[1]	Yes	NP[2]	[12]
Association graphs	Yes	Yes	No	Yes	NP	[14, 23]
Graph edition	Yes	Yes	Yes	Yes	NP	[7, 21, 36]
Random graphs	Yes	Yes	Yes	Yes	NP	[25, 38]
Probabilistic relaxation	Yes	Yes	Yes	No	P	[5, 8, 11, 37]
Neural networks	Yes	Yes	Yes	No	P	[16, 29, 28]
Genetic algorithms	Yes	Yes	Yes	No	P	[6, 9, 15]
Eigendecomposition	Yes	No	No[3]	Yes	P	[33]
Linear programming	Yes	No	No	Yes	P	[2]
Indexed search	Yes	Yes	No	Yes	P[4]	[4, 27]

[1] In some cases (e.g. [12]).
[2] If backtracking follows relaxation.
[3] Although is able to find error-tolerant graph isomorphism between close graphs.
[4] Although the compilation of the database is NP.

Table: In Terms of Their Computational Complexity and the Ability to Perform an Inexact Matching, [Lladós 2001].

Graph Edit Distance (ED)

The minimum amount of distortion that is needed to transform G_1 into G_2

- Distortions s_i: deletions, insertions, substitutions of nodes and edges.
- Edit path $S = s_1, ..., s_n$: A sequence of edit operations that transforms G_1 into G_2.
- Cost functions: Measuring the strength of a given distortion.
- Edit distance $d(G_1, G_2)$: Minimum cost edit path between two graphs.

Problem of Edit Distance: NP complete

- Explore the space of all possible mappings of the nodes and edges of G_1 to the nodes and edges of G_2.
- Edit Distance computation also has a worst case exponential complexity which prevents its use in large datasets.

Graph Edit Distance (ED)

The minimum amount of distortion that is needed to transform G_1 into G_2

- Distortions s_i: deletions, insertions, substitutions of nodes and edges.
- Edit path $S = s_1, ..., s_n$: A sequence of edit operations that transforms G_1 into G_2.
- Cost functions: Measuring the strength of a given distortion.
- Edit distance $d(G_1, G_2)$: Minimum cost edit path between two graphs.

Problem of Edit Distance: NP complete

- Explore the space of all possible mappings of the nodes and edges of G_1 to the nodes and edges of G_2.
- Edit Distance computation also has a worst case exponential complexity which prevents its use in large datasets.

Graph Edit Distance (ED)

The minimum amount of distortion that is needed to transform G_1 into G_2

- Distortions s_i: deletions, insertions, substitutions of nodes and edges.
- Edit path $S = s_1, ..., s_n$: A sequence of edit operations that transforms G_1 into G_2.
- Cost functions: Measuring the strength of a given distortion.
- Edit distance $d(G_1, G_2)$: Minimum cost edit path between two graphs.

Problem of Edit Distance: NP complete

- Explore the space of all possible mappings of the nodes and edges of G_1 to the nodes and edges of G_2.
- Edit Distance computation also has a worst case exponential complexity which prevents its use in large datasets.

Graph Edit Distance (ED)

The minimum amount of distortion that is needed to transform G_1 into G_2

- Distortions s_i: deletions, insertions, substitutions of nodes and edges.
- Edit path $S = s_1, ..., s_n$: A sequence of edit operations that transforms G_1 into G_2.
- Cost functions: Measuring the strength of a given distortion.
- Edit distance $d(G_1, G_2)$: Minimum cost edit path between two graphs.

Problem of Edit Distance: NP complete

- Explore the space of all possible mappings of the nodes and edges of G_1 to the nodes and edges of G_2.
- Edit Distance computation also has a worst case exponential complexity which prevents its use in large datasets.

Graph Edit Distance (ED)

The minimum amount of distortion that is needed to transform G_1 into G_2

- Distortions s_i: deletions, insertions, substitutions of nodes and edges.
- Edit path $S = s_1, ..., s_n$: A sequence of edit operations that transforms G_1 into G_2.
- Cost functions: Measuring the strength of a given distortion.
- Edit distance $d(G_1, G_2)$: Minimum cost edit path between two graphs.

Problem of Edit Distance: NP complete

- Explore the space of all possible mappings of the nodes and edges of G_1 to the nodes and edges of G_2.
- Edit Distance computation also has a worst case exponential complexity which prevents its use in large datasets.

Graph Edit Distance (ED)

The minimum amount of distortion that is needed to transform G_1 into G_2

- Distortions s_i: deletions, insertions, substitutions of nodes and edges.
- Edit path $S = s_1, ..., s_n$: A sequence of edit operations that transforms G_1 into G_2.
- Cost functions: Measuring the strength of a given distortion.
- Edit distance $d(G_1, G_2)$: Minimum cost edit path between two graphs.

Problem of Edit Distance: NP complete

- Explore the space of all possible mappings of the nodes and edges of G_1 to the nodes and edges of G_2.
- Edit Distance computation also has a worst case exponential complexity which prevents its use in large datasets.

Approximation to Graph Edit Distance (ED)

Different types of approximations were proposed in [Hidovic 2004].

- Vector space embedding of graphs [Lopresti 2003], [Bunke 2010].
- Spectral graph theory [Robles-Kelly 2005].
- Probabilistic methods [Myers 2000].
- Combinatorial optimization [Gold 1996], [Shokoufandeh 2006], [Riesen 2009],[Jouili 2009].

Approximation to Graph Edit Distance (ED)

Different types of approximations were proposed in [Hidovic 2004].

- Vector space embedding of graphs [Lopresti 2003], [Bunke 2010].
- Spectral graph theory [Robles-Kelly 2005].
- Probabilistic methods [Myers 2000].
- Combinatorial optimization [Gold 1996], [Shokoufandeh 2006], [Riesen 2009],[Jouili 2009].

Approximation to Graph Edit Distance (ED)

Different types of approximations were proposed in [Hidovic 2004].

- Vector space embedding of graphs [Lopresti 2003], [Bunke 2010].
- Spectral graph theory [Robles-Kelly 2005].
- Probabilistic methods [Myers 2000].
- Combinatorial optimization [Gold 1996], [Shokoufandeh 2006], [Riesen 2009],[Jouili 2009].

Approximation to Graph Edit Distance (ED)

Different types of approximations were proposed in [Hidovic 2004].

- Vector space embedding of graphs [Lopresti 2003], [Bunke 2010].
- Spectral graph theory [Robles-Kelly 2005].
- Probabilistic methods [Myers 2000].
- Combinatorial optimization [Gold 1996], [Shokoufandeh 2006], [Riesen 2009],[Jouili 2009].

Graph comparison through combinatorial optimization

Basic idea:

- Methods are based on an optimization procedure **mapping local substructures**
- Any node(u_n) from G_1 can be assigned to any node(v_m) of G_2,
- Incurring some **cost** that depends on the u_n-v_m assignment.
- It is required to map all nodes in such a way that the total cost of the assignment is **minimized**.

Cost matrix representation (C):

- C_{ij} correspond to the costs of assigning the i^{th} node of G_1 to the j^{th} node of G_2.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Graph comparison through combinatorial optimization

Basic idea:

- Methods are based on an optimization procedure **mapping local substructures**
- Any node(u_n) from G_1 can be assigned to any node(v_m) of G_2,
- Incurring some **cost** that depends on the u_n-v_m assignment.
- It is required to map all nodes in such a way that the total cost of the assignment is **minimized**.

Cost matrix representation (C):

- C_{ij} correspond to the costs of assigning the i^{th} node of G_1 to the j^{th} node of G_2.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Graph comparison through combinatorial optimization

Basic idea:

- Methods are based on an optimization procedure **mapping local substructures**
- Any node(u_n) from G_1 can be assigned to any node(v_m) of G_2,
- Incurring some **cost** that depends on the u_n-v_m assignment.
- It is required to map all nodes in such a way that the total cost of the assignment is **minimized**.

Cost matrix representation (C):

- C_{ij} correspond to the costs of assigning the i^{th} node of G_1 to the j^{th} node of G_2.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Graph comparison through combinatorial optimization

Basic idea:

- Methods are based on an optimization procedure **mapping local substructures**
- Any node(u_n) from G_1 can be assigned to any node(v_m) of G_2,
- Incurring some **cost** that depends on the u_n-v_m assignment.
- It is required to map all nodes in such a way that the total cost of the assignment is **minimized**.

Cost matrix representation (C):

- C_{ij} correspond to the costs of assigning the i^{th} node of G_1 to the j^{th} node of G_2.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Graph comparison through combinatorial optimization

Basic idea:

- Methods are based on an optimization procedure **mapping local substructures**
- Any node(u_n) from G_1 can be assigned to any node(v_m) of G_2,
- Incurring some **cost** that depends on the u_n-v_m assignment.
- It is required to map all nodes in such a way that the total cost of the assignment is **minimized**.

Cost matrix representation (C):

- C_{ij} correspond to the costs of assigning the i^{th} node of G_1 to the j^{th} node of G_2.

$$C = \begin{array}{cc} \text{V1} & \\ \left|\begin{matrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{matrix}\right| & \text{V2} \end{array}$$

Combinatorial optimization: Comparative study

	Node signature	Distance
[Gold 1996]	Node degree+Label	*
[Shokoufandeh 2006]	Eigen vector	L2
[Riesen 2009]	(1)Node+(2)Edge	Edit cost

*: Depends on the graph attribute type.

Our proposal

- A generalization of prior works
- Where local substructures are represented as graphs
- Where the cost function $c(i, j)$ is a graph distance

Our proposal

- A generalization of prior works
- Where local substructures are represented as graphs
- Where the cost function $c(i, j)$ is a graph distance

Our proposal

- A generalization of prior works
- Where local substructures are represented as graphs
- Where the cost function $c(i, j)$ is a graph distance

Our proposal

- A generalization of prior works
- Where local substructures are represented as graphs
- Where the cost function $c(i, j)$ is a graph distance

A graph matching method based on subgraph assignments

Overview

- A distance between graph
- Subgraph decomposition
- Optimization algorithm

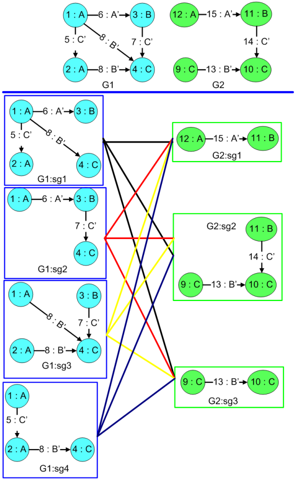


Figure: Subgraph matching: A bipartite graph

Overview

- A distance between graph
- Subgraph decomposition
- Optimization algorithm

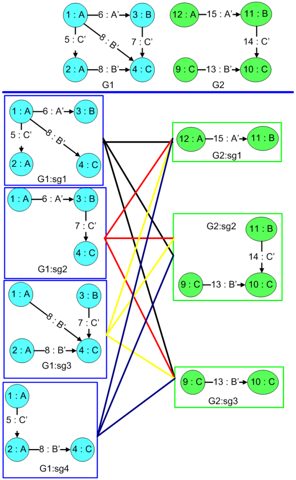


Figure: Subgraph matching: A bipartite graph

Overview

- A distance between graph
- Subgraph decomposition
- Optimization algorithm

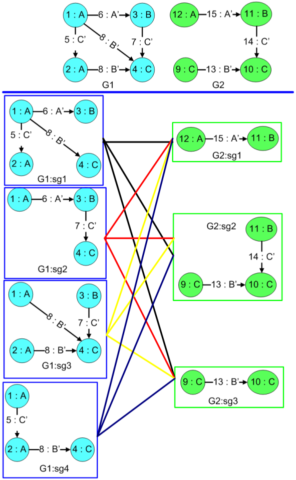


Figure: Subgraph matching: A bipartite graph

Graph decomposition

A subgraph (*sg*):

- A structure gathering the edges and their corresponding ending vertices from a root vertex.

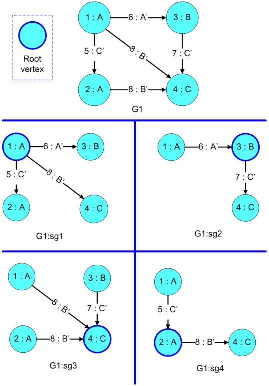


Figure: Decomposition into subgraph world

Matrix representation I

- The cost matrix contains the distances between every pair of subgraphs from G_1 and G_2.
  - What's the best (minimum-cost) way to assign the subgraphs?
- Assignment problem solved by the Hungarian method [Kuhn 1955]
- The cost of the minimum-weight subgraph matching :
  - SubGraph Matching Distance $SGMD(G_1, G_2)$

Example of possible variations of $SGMD$:

- $SGMD_{ED}$: Based on edit distance.
- $SGMD_{GP}$: Based on graph probing.

$$C = \begin{vmatrix} c_{1,1} & \dots & \dots & c_{1,m} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ c_{n,1} & \dots & \dots & c_{n,m} \end{vmatrix}$$

Matrix representation I

- The cost matrix contains the distances between every pair of subgraphs from G_1 and G_2.
  - What's the best (minimum-cost) way to assign the subgraphs?
- Assignment problem solved by the Hungarian method [Kuhn 1955]
- The cost of the minimum-weight subgraph matching :
  - SubGraph Matching Distance $SGMD(G_1, G_2)$

Example of possible variations of $SGMD$:

- $SGMD_{ED}$: Based on edit distance.
- $SGMD_{GP}$: Based on graph probing.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Matrix representation I

- The cost matrix contains the distances between every pair of subgraphs from G_1 and G_2.
  - What's the best (minimum-cost) way to assign the subgraphs?
- Assignment problem solved by the Hungarian method [Kuhn 1955]
- The cost of the minimum-weight subgraph matching :
  - SubGraph Matching Distance $SGMD(G_1, G_2)$

Example of possible variations of $SGMD$:

- $SGMD_{ED}$: Based on edit distance.
- $SGMD_{GP}$: Based on graph probing.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Matrix representation I

- The cost matrix contains the distances between every pair of subgraphs from G_1 and G_2.
  - What's the best (minimum-cost) way to assign the subgraphs?
- Assignment problem solved by the Hungarian method [Kuhn 1955]
- The cost of the minimum-weight subgraph matching :
  - SubGraph Matching Distance $SGMD(G_1, G_2)$

Example of possible variations of $SGMD$:

- $SGMD_{ED}$: Based on edit distance.
- $SGMD_{GP}$: Based on graph probing.

$$C = \begin{vmatrix} c_{1,1} & \cdots & \cdots & c_{1,m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{n,1} & \cdots & \cdots & c_{n,m} \end{vmatrix}$$

Theoretical discussion

- Our proposal is a pseudo metric
  - Positive
  - Symmetric
  - Triangle inequality
- $SGMD_{ED}$ is a lower bound for the edit distance
  - $\forall G_1, G_2 : \frac{SGMD_{ED}(G_1, G_2)}{\max(|G_1|, |G_2|)} \leq ED(G_1, G_2)$

Theoretical discussion

- Our proposal is a pseudo metric
  - Positive
  - Symmetric
  - Triangle inequality
- $SGMD_{ED}$ is a lower bound for the edit distance
  - $\forall G_1, G_2 : \frac{SGMD_{ED}(G_1, G_2)}{\max(|G_1|, |G_2|)} \leq ED(G_1, G_2)$

Theoretical discussion

- Our proposal is a pseudo metric
  - Positive
  - Symmetric
  - Triangle inequality
- $SGMD_{ED}$ is a lower bound for the edit distance
  - $\forall G_1, G_2 : \frac{SGMD_{ED}(G_1, G_2)}{\max(|G_1|, |G_2|)} \leq ED(G_1, G_2)$

Theoretical discussion

- Our proposal is a pseudo metric
  - Positive
  - Symmetric
  - Triangle inequality
- $SGMD_{ED}$ is a lower bound for the edit distance
  - $\forall G_1, G_2 : \frac{SGMD_{ED}(G_1, G_2)}{\max(|G_1|, |G_2|)} \leq ED(G_1, G_2)$

Theoretical discussion

- Our proposal is a pseudo metric
  - Positive
  - Symmetric
  - Triangle inequality
- $SGMD_{ED}$ is a lower bound for the edit distance
  - $\forall G_1, G_2 : \frac{SGMD_{ED}(G_1, G_2)}{\max(|G_1|, |G_2|)} \leq ED(G_1, G_2)$

Theoretical discussion

- Our proposal is a pseudo metric
  - Positive
  - Symmetric
  - Triangle inequality
- $SGMD_{ED}$ is a lower bound for the edit distance
  - $\forall G_1, G_2 : \dfrac{SGMD_{ED}(G_1, G_2)}{\max(|G_1|, |G_2|)} \leq ED(G_1, G_2)$

Experiments

Hypothesis:

- The more accurate the distance induced by graph matching is, the better the matching is.

The question turns into a graph distance comparison:

- Correlation
- Classification

Experiments

Hypothesis:

- The more accurate the distance induced by graph matching is, the better the matching is.

The question turns into a graph distance comparison:

- Correlation
- Classification

Experiments

Hypothesis:

- The more accurate the distance induced by graph matching is, the better the matching is.

The question turns into a graph distance comparison:

- Correlation
- Classification

Databases

- IAM Graph Database Repository (Standardized graph data sets for benchmarking).
- Synthetic data set (Randomly generated for scalability testing).
- Home-made data sets (Domain-dependent applications).

Table: Characteristics of the four data sets used in our computational experiments

	Base A	Base B	Base C	Base D
Number of classes (N)	50	10	32	15
| *Training* |	14128	114	9600	5062
| *Validation* |	14101	56	3200	1688
Average number of nodes	12.03	5.56	8.84	4.7
Average number of edges	9.86	11.71	10.15	3.6
Average degree of nodes	1.63	4.21	1.15	1.3

Databases

- IAM Graph Database Repository (Standardized graph data sets for benchmarking).
- Synthetic data set (Randomly generated for scalability testing).
- Home-made data sets (Domain-dependent applications).

Table: Characteristics of the four data sets used in our computational experiments

	Base A	Base B	Base C	Base D
Number of classes (N)	50	10	32	15
| *Training* |	14128	114	9600	5062
| *Validation* |	14101	56	3200	1688
Average number of nodes	12.03	5.56	8.84	4.7
Average number of edges	9.86	11.71	10.15	3.6
Average degree of nodes	1.63	4.21	1.15	1.3

Databases

- IAM Graph Database Repository (Standardized graph data sets for benchmarking).
- Synthetic data set (Randomly generated for scalability testing).
- Home-made data sets (Domain-dependent applications).

Table: Characteristics of the four data sets used in our computational experiments

	Base A	Base B	Base C	Base D
Number of classes (N)	50	10	32	15
| *Training* |	14128	114	9600	5062
| *Validation* |	14101	56	3200	1688
Average number of nodes	12.03	5.56	8.84	4.7
Average number of edges	9.86	11.71	10.15	3.6
Average degree of nodes	1.63	4.21	1.15	1.3

Protocol

- Correlation between ED and suboptimal distances:
  - Rank correlation: Kendall correlation
  - Distance correlation: Pearson correlation
- Classification stage
  - 1-NN classifier

Protocol

- Correlation between ED and suboptimal distances:
  - Rank correlation: Kendall correlation
  - Distance correlation: Pearson correlation
- Classification stage
  - 1-NN classifier

Protocol

- Correlation between ED and suboptimal distances:
  - Rank correlation: Kendall correlation
  - Distance correlation: Pearson correlation
- Classification stage
  - 1-NN classifier

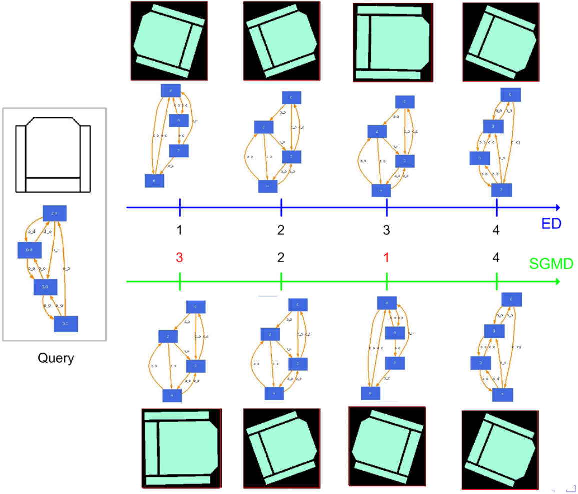
Protocol

- Correlation between ED and suboptimal distances:
  - Rank correlation: Kendall correlation
  - Distance correlation: Pearson correlation
- Classification stage
  - 1-NN classifier

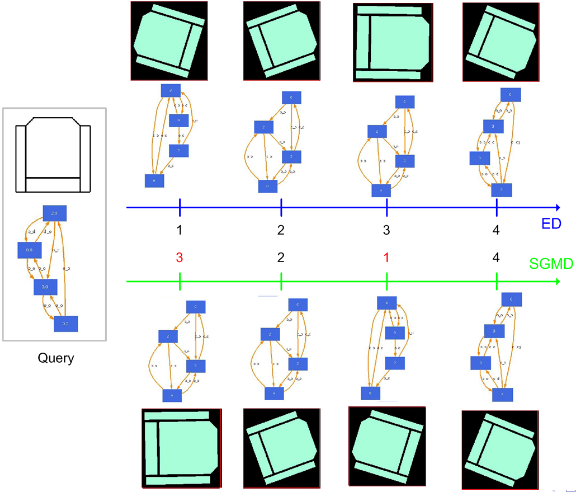
Protocol

- Correlation between ED and suboptimal distances:
  - Rank correlation: Kendall correlation
  - Distance correlation: Pearson correlation
- Classification stage
  - 1-NN classifier

Rank relationship with edit distance I



Rank relationship with edit distance I



	1	2
	1	1
	2	2
	3	4
	4	3
τ	1	$\frac{2}{3}$

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
    - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
  - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
  - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
  - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
  - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
  - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Rank relationship with edit distance II

$SGMD_{ED}$ vs ED

- $M = 1200$ queries.
- Top k responses to each query (k=30)
- A null hypothesis of independence($H0$) between the two responses
- Ranks are observed as ordered categorical variables
- Kendall correlation coefficient(τ) is computed for each query pair ($SGMD_{ED}$ vs ED)
- From the 1200 tests, only 124 have a p-value greater than 0.05
  - 124 queries did not pass the Kendall's test
- $H0$ can be rejected in 89.67% cases, with a risk of 5%.

Classification stage

The standard nearest-neighbor $(1 - NN)$ classification rule assigns x to the class of the *most similar* graph in a set of labeled training data.

Table: Classification rate according to the graph distance in use

Method	Base A	Base B	Base C	Base D
$ED(\%)$	—	92.86	—	**82.10**
$SGMD_{ED}(\%)$	**88.54**	**94.64**	**99.54**	80.86
$SGMD_{GP}(\%)$	88.48	94.64	99.21	78.79
$GP(\%)$	57.01	92.86	98.33	59.89
$NMD(\%)$	29.49	89.28	88.75	36.96

Time complexity

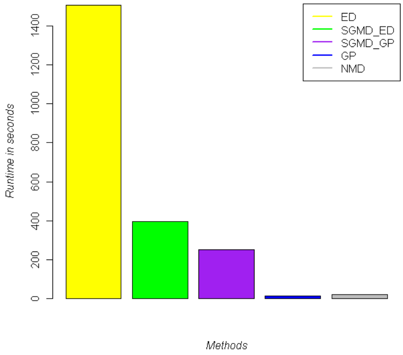


Figure: Time complexity

Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

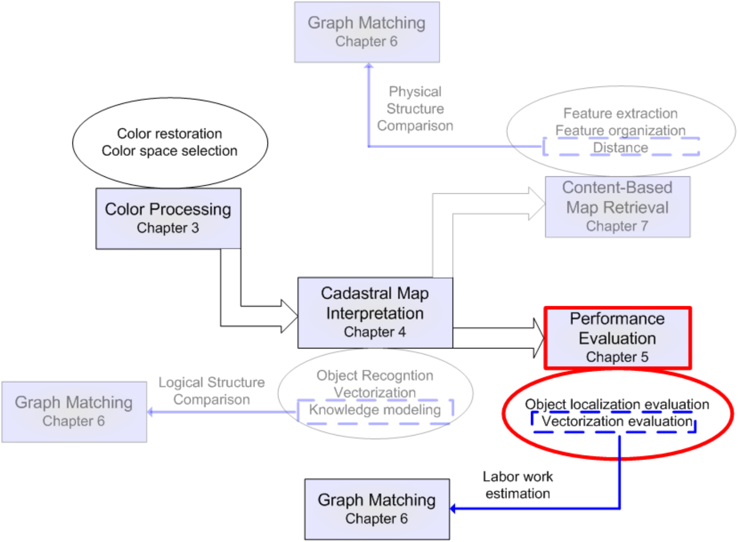
Bottom lines

- Graph matching algorithm
- Graph distance
- Polynomial time complexity ($O(n^3)$)
- Lower bound relation with the edit distance
- Rank relation with edit distance
- 1-NN classifier is not negatively affected by using sub-optimal distances.
- Flexible distance with two meta-parameters:
  - Sub distance
  - Subgraph size

Transition

- We have presented some general applications of graph comparison
- Next slides are dedicated to the use of graph distances in a context of performance evaluation

Main steps


Graph Matching
Chapter 6
Physical
Structure
Comparison
Feature extraction
Feature organization
Distance
Color restoration
Color space selection
Color Processing
Chapter 3
Content-Based
Map Retrieval
Chapter 7
Cadastral Map
Interpretation
Chapter 4
Performance
Evaluation
Chapter 5
Graph Matching
Chapter 6
Logical Structure
Comparison
Object Recogntion
Vectorization
Knowledge modeling
Object localization evaluation
Vectorization evaluation
Labor work
estimation
Graph Matching
Chapter 6

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
  - To determine the weaknesses and strengths

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
  - To determine the weaknesses

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
  - To determine the weaknesses and strengths

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
  - To determine the weaknesses and strengths

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
    - To determine the weaknesses and strengths

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
    - To determine the weaknesses and strengths

Evaluation of Vectorized Documents

Overview:

- Performance Evaluation(PE) has become of first interest during the last years.
- A contest on this topic: Since GREC'95 and every two years

The goal:

- A need for standard protocols to compare and evaluate methods.
- To establish a solid knowledge of the state of the art
  - To determine the weaknesses and strengths

Around performance evaluation for R2V system I

Performance evaluation of vectorization and line detection has been reported by [Kong 1996], [Hori 1996], [Wenyin 1997] and [Chhabra 1998].

- A method for evaluating the recognition of dashed lines
- Hori and Doermann [Hori 1996], a measurement methodology for task-specific raster to vector conversion
- Wenyin and Dori [Wenyin 1997], a protocol for evaluating the recognition of straight and circular lines
- Phillips and Chhabra [Chhabra 1998], a methodology for evaluating graphics recognition systems operating on images that contain straight lines and text blocks

Around performance evaluation for R2V system I

Performance evaluation of vectorization and line detection has been reported by [Kong 1996], [Hori 1996], [Wenyin 1997] and [Chhabra 1998].

- A method for evaluating the recognition of dashed lines
- Hori and Doermann [Hori 1996], a measurement methodology for task-specific raster to vector conversion
- Wenyin and Dori [Wenyin 1997], a protocol for evaluating the recognition of straight and circular lines
- Phillips and Chhabra [Chhabra 1998], a methodology for evaluating graphics recognition systems operating on images that contain straight lines and text blocks

Around performance evaluation for R2V system I

Performance evaluation of vectorization and line detection has been reported by [Kong 1996], [Hori 1996], [Wenyin 1997] and [Chhabra 1998].

- A method for evaluating the recognition of dashed lines
- Hori and Doermann [Hori 1996], a measurement methodology for task-specific raster to vector conversion
- Wenyin and Dori [Wenyin 1997], a protocol for evaluating the recognition of straight and circular lines
- Phillips and Chhabra [Chhabra 1998], a methodology for evaluating graphics recognition systems operating on images that contain straight lines and text blocks

Around performance evaluation for R2V system I

Performance evaluation of vectorization and line detection has been reported by [Kong 1996], [Hori 1996], [Wenyin 1997] and [Chhabra 1998].

- A method for evaluating the recognition of dashed lines
- Hori and Doermann [Hori 1996], a measurement methodology for task-specific raster to vector conversion
- Wenyin and Dori [Wenyin 1997], a protocol for evaluating the recognition of straight and circular lines
- Phillips and Chhabra [Chhabra 1998], a methodology for evaluating graphics recognition systems operating on images that contain straight lines and text blocks

Around performance evaluation for R2V system I

Performance evaluation of vectorization and line detection has been reported by [Kong 1996], [Hori 1996], [Wenyin 1997] and [Chhabra 1998].

- A method for evaluating the recognition of dashed lines
- Hori and Doermann [Hori 1996], a measurement methodology for task-specific raster to vector conversion
- Wenyin and Dori [Wenyin 1997], a protocol for evaluating the recognition of straight and circular lines
- Phillips and Chhabra [Chhabra 1998], a methodology for evaluating graphics recognition systems operating on images that contain straight lines and text blocks

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Around performance evaluation for R2V system II

- All these methods are limited in their applicability to the ALPAGE project.
- All prior works focused on a lower level of consistency (arcs and segments) where we need an evaluation at polygon level.
- Modification of previous methods to a polygon entity is not trivial
  - A higher level requires a matching task when segments do not
- We propose an extension to polygon level of related approaches
- Evaluation of Vectorized Documents by means of Polygon Assignments and a Graph-Based Dissimilarity

Problem definition

Two issues about the evaluation of the:

1. Polygon detection
2. Polygon approximation

Problem definition: Polygon detection

- Given two sets of polygons, D_1 and D_2.
- Associated together with a weight function $C : D_1 \times D_2 \rightarrow \mathbb{R}$
- Find a mapping $f : D_1 \rightarrow D_2$ such that the cost function Eq. 1 is minimized

$$\sum_{p \in D_1} C(p, f(p)), \quad (1)$$

where p is a polygon

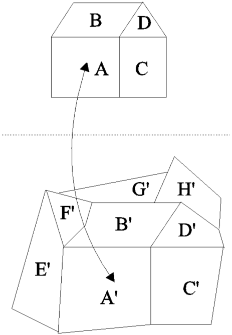
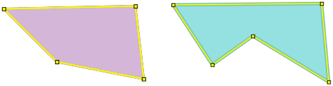


Figure: Polygon partitions. (up) D_1; (down) D_2

Problem definition: Polygon approximation

- Given two polygons P_1, P_2 with N and M points, respectively.
- The approximation error between P_1 and P_2 , $d(P_1, P_2)$.



(a) P_1 (b) P_2

Figure: Polygons to be compared.

Toward a proposition for evaluating polygon detection algorithms

Our proposal for assessing the quality of polygon detection system:

Two viewpoints:

1. Polygon location
2. Polygon approximation

A Local Evaluation of Vectorized Documents by means of Polygon Assignments and Matching

Toward a proposition for evaluating polygon detection algorithms

Our proposal for assessing the quality of polygon detection system:

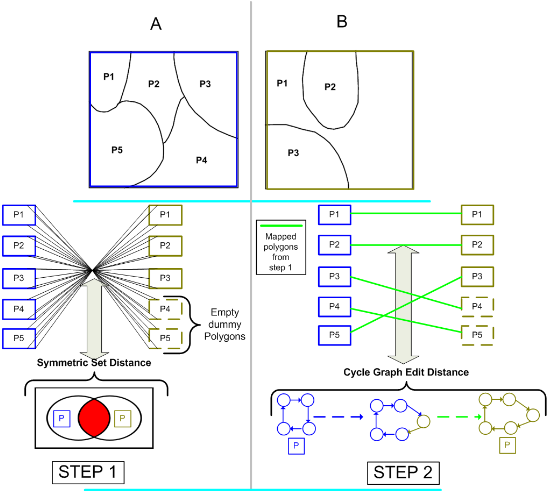
Two viewpoints:

1. Polygon location
2. Polygon approximation

A Local Evaluation of Vectorized Documents by means of Polygon Assignments and Matching

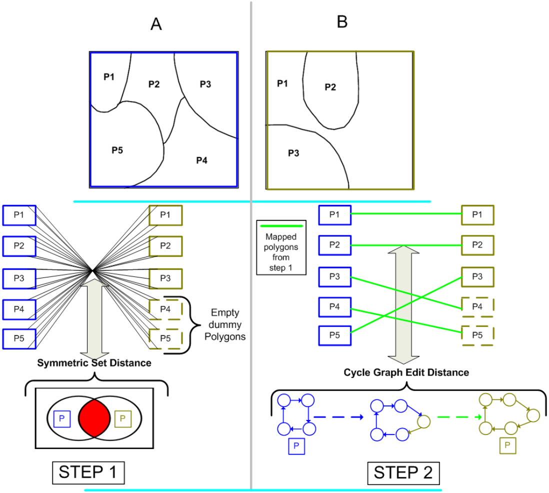
Overview

1. A bipartite graph weighed by the symmetric difference
   - To evaluate how well polygons are detected and located
2. A cycle graph edit distance applied to polygons
   - The correctness of the polygonal approximation (Vectorization precision).



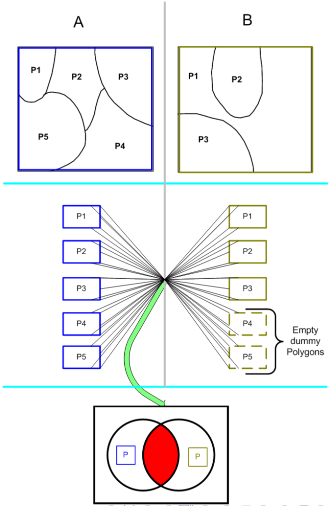
Overview

1. A bipartite graph weighed by the symmetric difference
   - To evaluate how well polygons are detected and located
2. A cycle graph edit distance applied to polygons
   - The correctness of the polygonal approximation (Vectorization precision).



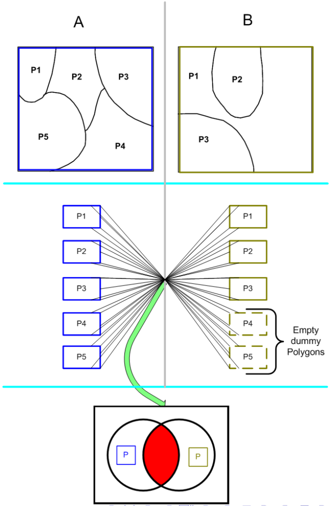
Step 1

1. $K : P \times P \to \mathbb{R}$
2. Optimization algorithm
   - What's the best (minimum-cost) way to assign the polygons?
3. Assignment problem solved by the Hungarian method [Kuhn 1955]
4. The cost of the minimum-weight polygon mapping :
   - Polygon Mapping Distance $PMD(D_1, D_2)$



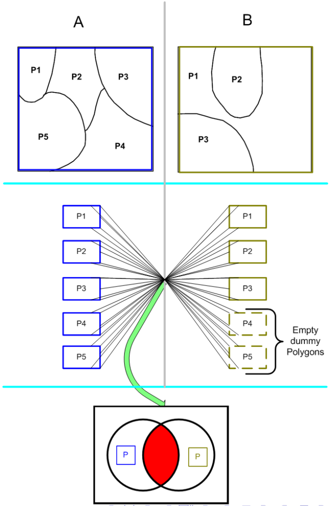
Step 1

1. $K : P \times P \rightarrow \mathbb{R}$
2. Optimization algorithm
   - What's the best (minimum-cost) way to assign the polygons?
3. Assignment problem solved by the Hungarian method [Kuhn 1955]
4. The cost of the minimum-weight polygon mapping :
   - Polygon Mapping Distance $PMD(D_1, D_2)$



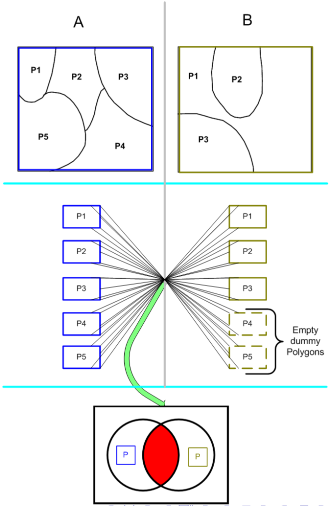
Step 1

1. $K : P \times P \to \mathbb{R}$
2. Optimization algorithm
   - What's the best (minimum-cost) way to assign the polygons?
3. Assignment problem solved by the Hungarian method [Kuhn 1955]
4. The cost of the minimum-weight polygon mapping :
   - Polygon Mapping Distance $PMD(D_1, D_2)$



Step 1

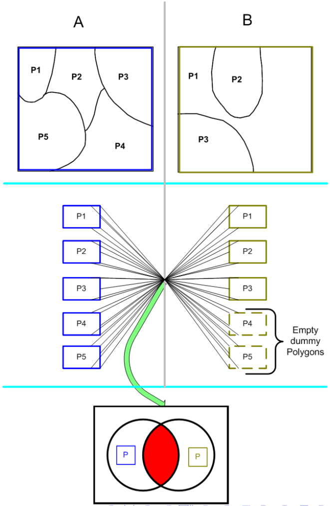
1. $K : P \times P \to \mathbb{R}$
2. Optimization algorithm
   - What's the best (minimum-cost) way to assign the polygons?
3. Assignment problem solved by the Hungarian method [Kuhn 1955]
4. The cost of the minimum-weight polygon mapping :
   - Polygon Mapping Distance $PMD(D_1, D_2)$



Step 1

Weaknesses:

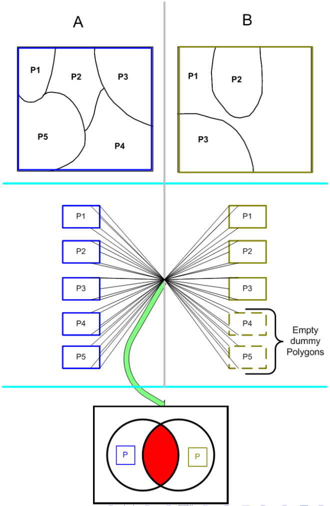
- K provides a location information.
- K does not take into account the labor work that has to be done to change a polygon from the CG to a correct polygon from the GT.
- An additional information needed.



Step 1

Weaknesses:

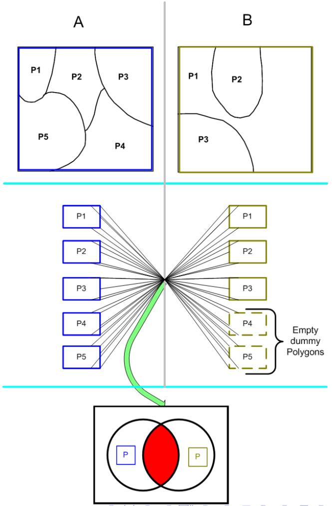
- K provides a location information.
- K does not take into account the labor work that has to be done to change a polygon from the CG to a correct polygon from the GT.
- An additional information needed.



Step 1

Weaknesses:

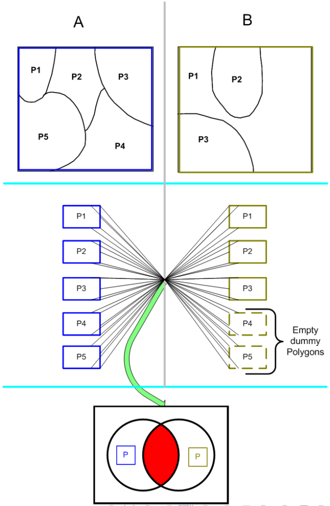
- K provides a location information.
- K does not take into account the labor work that has to be done to change a polygon from the CG to a correct polygon from the GT.
- An additional information needed.



Step 1

Weaknesses:

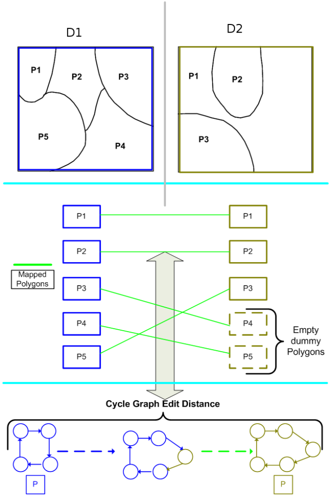
- K provides a location information.
- K does not take into account the labor work that has to be done to change a polygon from the CG to a correct polygon from the GT.
- An additional information needed.



Step 2

Labor work consideration:

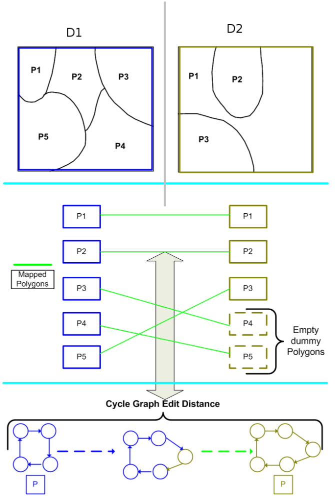
- To reveal how many edit operations have to be done to change a polygon into another according to some basic operations.
- Cycle Graph Edit Distance (CGED) for polygon comparison



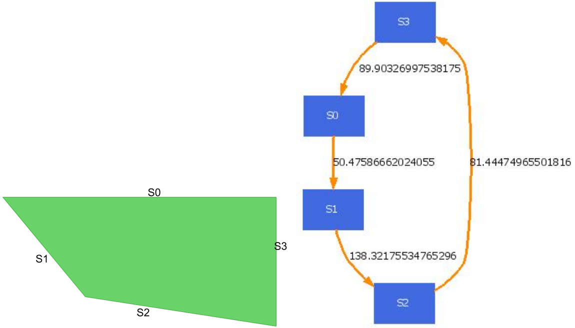
Step 2

Labor work consideration:

- To reveal how many edit operations have to be done to change a polygon into another according to some basic operations.
- Cycle Graph Edit Distance (CGED) for polygon comparison



From graph to polygon



(a) Polygon (b) Cycle Graph

Figure: From polygon to cycle graph

The problem turns into a graph comparison problem.

Graph comparison

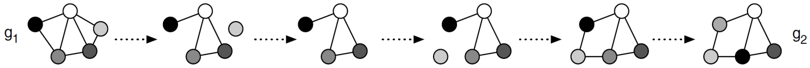


Figure: A possible edit path between graph g1 and g2 (node labels are represented by different shades of grey)[Riesen 2009]

The cost functions for attributed cycle graph matching are:

Table: Edit costs

	Node	Edge
Label Substitution	$\gamma((l_i^A) \rightarrow (l_j^B)) = \left\| \frac{l_i^A}{\|A\|} - \frac{l_j^B}{\|B\|} \right\|$	$\gamma((\Phi_i^A) \rightarrow (\Phi_j^B)) = \frac{\|\Phi_i^A - \Phi_j^B\|}{360}$
Addition	$\gamma(\lambda \rightarrow (l_j^B)) = \frac{l_j^B}{\|B\|}$	$\gamma(\lambda \rightarrow (\Phi_j^B)) = \frac{\|\Phi_j^B\|}{360}$
Deletion	$\gamma((l_i^A) \rightarrow \lambda) = \frac{l_i^A}{\|A\|}$	$\gamma((\Phi_i^A) \rightarrow \lambda) = \frac{\|\Phi_i^A\|}{360}$

Operation on polygons

Editing a vectorization with the basic operations are:

- Add
- Delete
- Move

Impact on the graph representation

- Through linear combinations
- It is possible to recreate the usages of a person modifying a vectorization

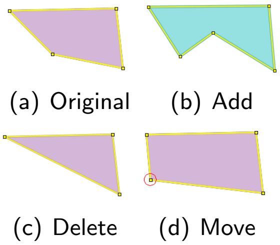


Figure: Basic edit operations applied to a polygon.

Operation on polygons

Editing a vectorization with the basic operations are:

- Add
- Delete
- Move

Impact on the graph representation:

- Through linear combinations
- It is possible to recreate the usages of a person modifying a vectorization

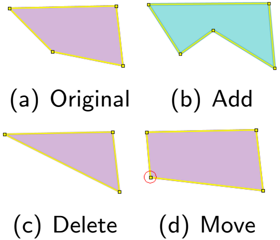


Figure: Basic edit operations applied to a polygon.

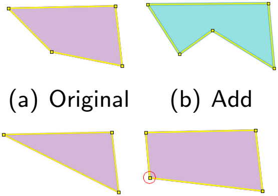
Operation on polygons

Editing a vectorization with the basic operations are:

- Add
- Delete
- Move

Impact on the graph representation

- Through linear combinations
- It is possible to recreate the usages of a person modifying a vectorization



(a) Original (b) Add (c) Delete (d) Move

Figure: Basic edit operations applied to a polygon.

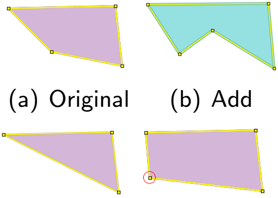
Operation on polygons

Editing a vectorization with the basic operations are:

- Add
- Delete
- Move

Impact on the graph representation

- Through linear combinations
- It is possible to recreate the usages of a person modifying a vectorization



(a) Original (b) Add

(c) Delete (d) Move

Figure: Basic edit operations applied to a polygon.

Operation on polygons

Editing a vectorization with the basic operations are:

- Add
- Delete
- Move

Impact on the graph representation

- Through linear combinations
- It is possible to recreate the usages of a person modifying a vectorization

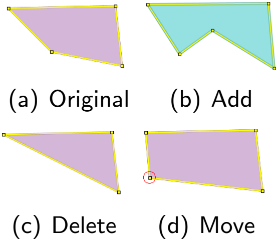


Figure: Basic edit operations applied to a polygon.

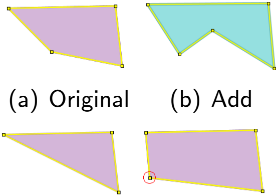
Operation on polygons

Editing a vectorization with the basic operations are:

- Add
- Delete
- Move

Impact on the graph representation

- Through linear combinations
- It is possible to recreate the usages of a person modifying a vectorization



(a) Original (b) Add

(c) Delete (d) Move

Figure: Basic edit operations applied to a polygon.

Impact on the graph representation

Edit operation for segment deletion

- Delete a segment = Delete a node and an edge into the graph
- We conclude:

$$deletions = \gamma((I_i^A) \to \lambda) + \gamma((\Phi_i^A) \to \lambda)$$

- Theses deletions create an orphan edge that must be reconnected
- Consequently we conclude:

$$substitution = \gamma((\Phi_i^A) \to (\Phi_j^B))$$

Impact on the graph representation

Edit operation for segment deletion

- Delete a segment = Delete a node and an edge into the graph
- We conclude:

$$deletions = \gamma((I_i^A) \rightarrow \lambda) + \gamma((\Phi_i^A) \rightarrow \lambda)$$

- Theses deletions create an orphan edge that must be reconnected
- Consequently we conclude:

$$substitution = \gamma((\Phi_i^A) \rightarrow (\Phi_j^B))$$

Impact on the graph representation

Edit operation for segment deletion

- Delete a segment = Delete a node and an edge into the graph
- We conclude:

$$deletions = \gamma((I_i^A) \to \lambda) + \gamma((\Phi_i^A) \to \lambda)$$

- Theses deletions create an orphan edge that must be reconnected
- Consequently we conclude:

$$substitution = \gamma((\Phi_i^A) \to (\Phi_j^B))$$

Impact on the graph representation

Edit operation for segment deletion

- Delete a segment = Delete a node and an edge into the graph
- We conclude:

$$deletions = \gamma((l_i^A) \rightarrow \lambda) + \gamma((\Phi_i^A) \rightarrow \lambda)$$

- Theses deletions create an orphan edge that must be reconnected
- Consequently we conclude:

$$substitution = \gamma((\Phi_i^A) \rightarrow (\Phi_j^B))$$

Impact on the graph representation

Edit operation for segment deletion

- Finally, the sequence of operations is:

$$\gamma(s_i \rightarrow \lambda) = deletions + substitution$$

- Simply, we swap formal expressions by their corresponding costs:

$$\gamma(s_i \rightarrow \lambda) = \frac{l_i^A}{|A|} + \frac{\Phi_i^A}{360} + \frac{|\Phi_i^A - \Phi_j^B|}{360}$$

Impact on the graph representation

Edit operation for segment deletion

- Finally, the sequence of operations is:

$$\gamma(s_i \to \lambda) = \mathit{deletions} + \mathit{substitution}$$

- Simply, we swap formal expressions by their corresponding costs:

$$\gamma(s_i \to \lambda) = \frac{l_i^A}{|A|} + \frac{\Phi_i^A}{360} + \frac{|\Phi_i^A - \Phi_j^B|}{360}$$

Experiments

- Database description
- Protocol definition
- Tests:
  - Polygon Matching Distance (PMD)
  - Matched Edit Distance (MED)
  - Cadastral parcel retrieval evaluation
- are PMD and MED representative of polygon deformations ?
  - Shape variation
  - Polygonal approximation modification

Experiments

- Database description
- Protocol definition
- Tests:
  - Polygon Matching Distance (PMD)
  - Matched Edit Distance (MED)
  - Cadastral parcel retrieval evaluation
- are PMD and MED representative of polygon deformations ?
  - Shape variation
  - Polygonal approximation modification

Experiments

- Database description
- Protocol definition
- Tests:
  - Polygon Matching Distance (PMD)
  - Matched Edit Distance (MED)
  - Cadastral parcel retrieval evaluation
- are PMD and MED representative of polygon deformations ?
  - Shape variation
  - Polygonal approximation modification

Experiments

- Database description
- Protocol definition
- Tests:
  - Polygon Matching Distance (PMD)
  - Matched Edit Distance (MED)
  - Cadastral parcel retrieval evaluation
- are PMD and MED representative of polygon deformations ?
  - Shape variation
  - Polygonal approximation modification

Databases I

- **Base A** Shape distortion: Derived from [Delalandre 2010] and [Dosch 2006]
    - To evaluate polygon detection methods

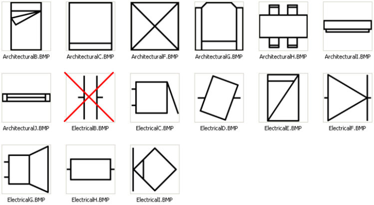
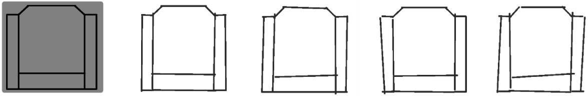
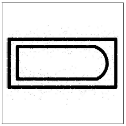


Figure: A sample among the seventy symbols used in our ranking test.

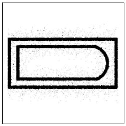


Databases II

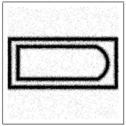
- **Base B** Binary degradation: From the data set provided by the GREC'03 contest.
    - The higher is the noise level the higher are the distortions on the polygonal approximation.



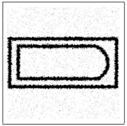
(a) Noise level 1.



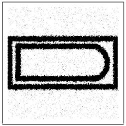
(b) Noise level 2.



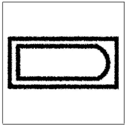
(c) Noise level 3.



(d) Noise level 4.

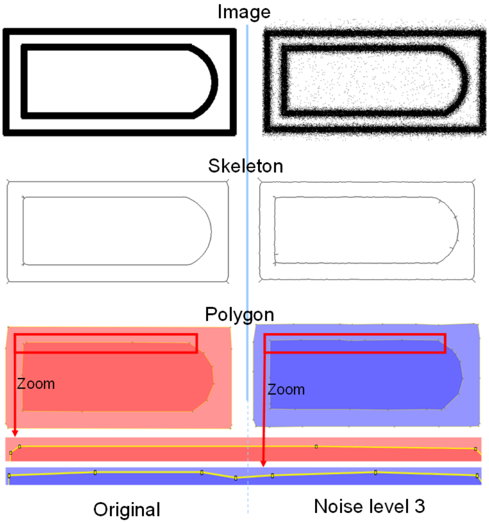


(e) Noise level 5.

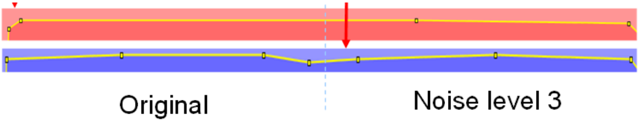


(f) Noise level 6.

Databases III

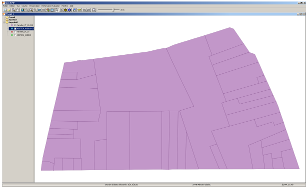


Databases III

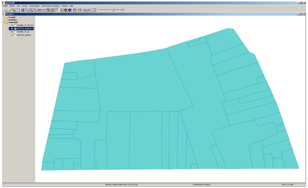


Databases IV

- **Base C** Cadastral map collection from ALPAGE project
- Computer generated elements (CG).
- Manually vectorized references (GT).



(a) GT



(b) CG

Figure: Two vectorizations to be mapped ($|D_{CG}| = 46$ $|D_{GT}| = 40$).

Protocol

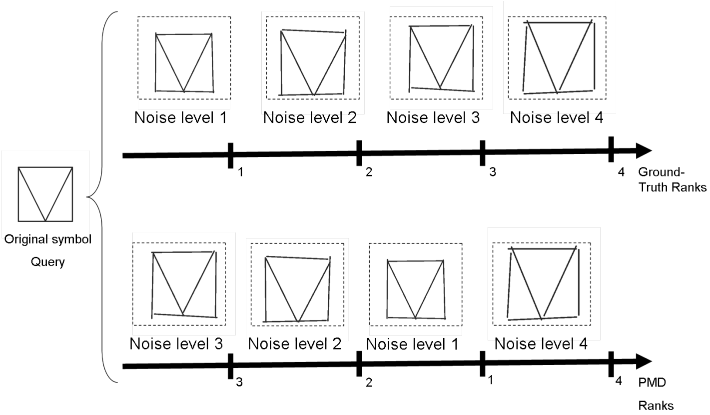


Figure: Ranking explanation. Ranks 3 and 1 were swapped by PMD

Ranking

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
τ	0.0000	0.6000	0.8000	0.7029	0.8000	1.0000

Table: Summary of Kendall correlation (τ). PMD *vs* ground-truth

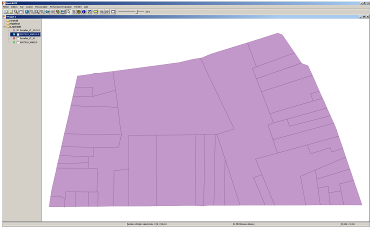
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
τ	0.3333	0.6190	0.7143	0.7107	0.8095	1.0000

Table: Summary of Kendall correlation (τ). MED *vs* ground-truth

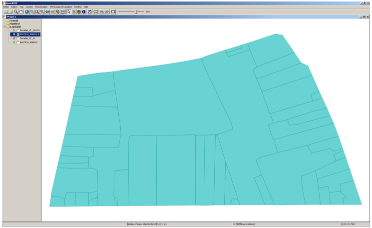
Application to the evaluation of parcel detection I

A visual dissimilarity measure of local anomalies:

- Comparing maps two by two.
- It facilitates the spotting of errors
- A visual signs are worth a thousand words



(a) GT



(b) CG

Figure: Two vectorizations to be mapped ($|D_{CG}| = 46$ $|D_{GT}| = 40$).

Application to the evaluation of parcel detection II

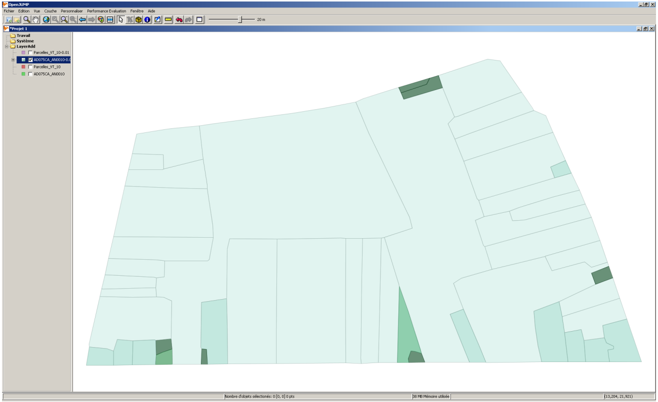


Figure: Local dissimilarities between the two maps. The lighter the better.

Summary I

- A protocol for performance evaluation of polygon detection algorithms.
- Our protocol is positioned as an extension of prior works, an extension at polygon level.
- Our contribution is two-fold
  - An object mapping algorithm to roughly locate errors within the drawing.
  - A cycle graph matching distance that depicts the accuracy of the polygonal approximation.

Summary I

- A protocol for performance evaluation of polygon detection algorithms.
- Our protocol is positioned as an extension of prior works, an extension at polygon level.
- Our contribution is two-fold
  - An object mapping algorithm to roughly locate errors within the drawing.
  - A cycle graph matching distance that depicts the accuracy of the polygonal approximation.

Summary I

- A protocol for performance evaluation of polygon detection algorithms.
- Our protocol is positioned as an extension of prior works, an extension at polygon level.
- Our contribution is two-fold
  - An object mapping algorithm to roughly locate errors within the drawing.
  - A cycle graph matching distance that depicts the accuracy of the polygonal approximation.

Summary II

- Both contributions were theoretically defined and adapted to the PE of polygonized documents.
  - A set distance for the polygon matching distance (PMD)
  - Dedicated edit costs for the graph matching method (MED)
- The behavior of our set of indices was analyzed when increasing image degradation.

Summary II

- Both contributions were theoretically defined and adapted to the PE of polygonized documents.
  - A set distance for the polygon matching distance (PMD)
  - Dedicated edit costs for the graph matching method (MED)
- The behavior of our set of indices was analyzed when increasing image degradation.

Summary II

- Both contributions were theoretically defined and adapted to the PE of polygonized documents.
  - A set distance for the polygon matching distance (PMD)
  - Dedicated edit costs for the graph matching method (MED)
- The behavior of our set of indices was analyzed when increasing image degradation.

Summary II

- Both contributions were theoretically defined and adapted to the PE of polygonized documents.
  - A set distance for the polygon matching distance (PMD)
  - Dedicated edit costs for the graph matching method (MED)
- The behavior of our set of indices was analyzed when increasing image degradation.

Conclusion and perspective

- Algorithms and methodologies designed in the ALPAGE project context.
- A team work: Many interactions with multidisciplinary people.
- Key: communication and listening to each other

Conclusion and perspective

- Algorithms and methodologies designed in the ALPAGE project context.
- A team work: Many interactions with multidisciplinary people.
- Key: communication and listening to each other

Conclusion and perspective

- Algorithms and methodologies designed in the ALPAGE project context.
- A team work: Many interactions with multidisciplinary people.
- Key: communication and listening to each other

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

To take the stock on our work

Summary:

- Color image analysis
- A modeling stage
- Domain-object extraction
- Graph-based representation
- Graph comparison
- Vectorization evaluation
- Vector to be inserted in a Geographic Information System.

Perspective

Near future:

- **Graph**: To check out the influence of subgraph matching with different depths $(1,2,...,|G|)$.
    - Will it increase the accuracy in classification ?
    - What about time consumption ?
- **Performance Evaluation**: To extend method to more complex objects than polygons.
    - To extend the concept to connected segments around the root polygon in order to constitute piece of symbols.
    - To change the scope of our performance evaluation tool to the direction of object spotting.

Related work:

- According this formulation, we are close to the work of [Lladós 2001].
    - J. Lladós et al, "Symbol Recognition by Error-Tolerant Subgraph Matching between Region Adjacency Graphs", *IEEE TPAMI*, vol. 23, 2001, pp. 1137-1143.

Perspective

Near future:

- **Graph**: To check out the influence of subgraph matching with different depths $(1,2,\ldots,|G|)$.
  - Will it increase the accuracy in classification ?
  - What about time consumption ?
- **Performance Evaluation**: To extend method to more complex objects than polygons.
  - To extend the concept to connected segments around the root polygon in order to constitute piece of symbols.
  - To change the scope of our performance evaluation tool to the direction of object spotting.

Related work:

- According this formulation, we are close to the work of [Lladós 2001].
  - J. Lladós et al, "Symbol Recognition by Error-Tolerant Subgraph Matching between Region Adjacency Graphs", *IEEE TPAMI*, vol. 23, 2001, pp. 1137-1143.

Perspective

Near future:

- **Graph**: To check out the influence of subgraph matching with different depths $(1,2,...,|G|)$.
  - Will it increase the accuracy in classification ?
  - What about time consumption ?
- **Performance Evaluation**: To extend method to more complex objects than polygons.
  - To extend the concept to connected segments around the root polygon in order to constitute piece of symbols.
  - To change the scope of our performance evaluation tool to the direction of object spotting.

Related work:

- According this formulation, we are close to the work of [Lladós 2001].
  - J. Lladós et al, "Symbol Recognition by Error-Tolerant Subgraph Matching between Region Adjacency Graphs" *IEEE TPAMI*, vol. 23, 2001, pp. 1137-1143.

Food for thought

Long-term future: **Semantic**

- Graph Matching for Model or Ontology Comparison
- Image Processing Driven by Knowledge

Food for thought

Long-term future: **Semantic**

- Graph Matching for Model or Ontology Comparison
- Image Processing Driven by Knowledge

Key contributions

- *Graph classification*: Published in **PRL**:
  - R. Raveaux J.-C. Burie and J.-M. Ogier. "A graph matching method and a graph matching distance based on subgraph assignments", *Pattern Recognition Letters*, 2009.
- *Performance Evaluation*: Accepted in **IJDAR**:
  - R. Raveaux, J-C Burie and J-M Ogier. "A Local Evaluation of Vectorized Documents by means of Polygon Assignments and Matching",*International Journal on Document Analysis and Recognition*, 2010.
- *Graph mining*: On the way, round 2 in **CVIU**:
  - R. Raveaux, S. Adam, P. Héroux, E. Trupin. "Learning Graph Prototypes for Shape Recognition", *Computer Vision and Image Understanding* .

Thank you for your attention

Some links:
ALPAGE: http://lamop.univ-paris1.fr/alpage/
Software: http://alpage-l3i.univ-lr.fr/
Contact: http://romain.raveaux.free.fr/

H Bunke.
On a relation between graph edit distance and maximum common subgraph.
Pattern Recognition Letters, vol. 18, no. 9, pages 689–694, 1997.

Horst Bunke and Kaspar Riesen.
Improving vector space embedding of graphs through feature selection algorithms.
Pattern Recognition, vol. In Press, Corrected Proof, pages –, 2010.

A Chhabra and I Phillips.
The Second International Graphics Recognition Contest - Raster to Vector Conversion: A Report.
Graphics Recognition: Algorithms and Systems, Lecture Notes in Computer Science, Springer, vol. 1389, 1998.

Mathieu Delalandre, Ernest Valveny, Tony Pridmore and Dimosthenis Karatzas.
Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems.
International Journal on Document Analysis and Recognition, page Online first, 2010.

Philippe Dosch and Ernest Valveny.
Report on the Second Symbol Recognition Contest, 2006.

Steven Gold and Anand Rangarajan.
Graph matching by graduated assignment.
IEEE transactions on pattern analysis and machine intelligence, pages 239–244, 1996.

Dzena Hidovic and Marcello Pelillo.
Metrics For Attributed Graphs Based On The Maximal Similarity Common Subgraph.

International Journal of Pattern Recognition and Artificial Intelligence, vol. 18, no. 3, pages 299–313, 2004.

O Hori and D S Doermann.
Quantitative measurement of the performance of raster-to-vector conversion algorithms.
Graphics recognition – methods and applications (Lecture Notes in Computer Science), vol. 1072, pages 57–68, 1996.

Salim Jouili and Salvatore Tabbone.
Graph Matching Based on Node Signatures.
In Graph-Based Representations in Pattern Recognition, pages 154–163, 2009.

B Kong, I T Phillips, R M Haralick, A Prasad and R Kasturi.
A benchmark: performance evaluation of dashed-line detection algorithms.

Graphics recognition – methods and applications (Lecture Notes in Computer Science), vol. 1072, pages 270–285, 1996.

H W Kuhn.
The Hungarian method for the assignment problem.
Naval Research Logistic Quarterly, vol. 2, pages 83–97, 1955.

J Lladós, E Marti and J J Villanueva.
Symbol Recognition by Error-Tolerant Subgraph Matching between Region Adjacency Graphs.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pages 1137–1143, 2001.

D Lopresti and G Wilfong.
A fast technique for comparing graph representations with applications to performance evaluation.
International Journal on Document Analysis and Recognition, vol. 6, no. 4, pages 219–229, 2003.

Richard Myers, Richard C Wilson and Edwin R Hancock.
Bayesian Graph Edit Distance.
IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 6, pages 628–635, 2000.

Kaspar Riesen and Horst Bunke.
Approximate graph edit distance computation by means of bipartite graph matching.
Image Vision Comput., vol. 27, no. 7, pages 950–959, 2009.

Antonio Robles-Kelly and Edwin R Hancock.
Graph Edit Distance from Spectral Seriation.
IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 3, pages 365–378, 2005.

Ali Shokoufandeh, Lars Bretzner, Diego Macrini, M Fatih Demirci, Clas Jönsson and Sven Dickinson.
The representation and matching of categorical shape.

Computer Vision and Image Understanding, vol. 103, no. 2, pages 139–154, 2006.

Liu Wenyin and Dov Dori.
A protocol for performance evaluation of line detection algorithms.
Machine Vision and Applications, vol. 9, no. 5, pages 240–250, 1997.