Cotutelle PhD thesis

Fuzzy Multilevel Graph Embedding for Recognition, Indexing and Retrieval of Graphic Document Images

presented by

Muhammad Muzzamil LUQMAN

mluqman@{univ-tours.fr, cvc.uab.es}

Friday, 2nd of March 2012

Directors of thesis

Dr. Jean-Yves RAMEL
Professor
University of Tours, France

Dr. Josep LLADOS
Professor
UAB, Spain

Co-supervisor

Dr. Thierry BROUARD
Assistant Professor
University of Tours, France

Higher Education Commission Pakistan

www.hec.gov.pk

❖ Problematic

- Lack of efficient computational tools for graph based structural pattern recognition

❖ Proposed solution

- Transform graphs into numeric feature vectors and exploit computational strengths of state of the art statistical pattern recognition

- Introduction
- Fuzzy Multilevel Graph Embedding (FMGE)
- Automatic indexing of graph repositories for graph retrieval and subgraph spotting
- Conclusions and future research challenges

❖ Introduction

- Structural and statistical pattern recognition
- Graph embedding
- State of the art on explicit graph embedding
- Limitations of existing methods

❖ Fuzzy Multilevel Graph Embedding (FMGE)

❖ Automatic indexing of graph repositories for graph retrieval and subgraph spotting

❖ Conclusions and future research challenges

	Pattern Recognition	
	Structural	**Statistical**
Data structure	symbolic data structure	numeric feature vector
Representational strength	Yes	No
Fixed dimensionality	No	Yes
Sensitivity to noise	Yes	No
Efficient computational tools	No	Yes

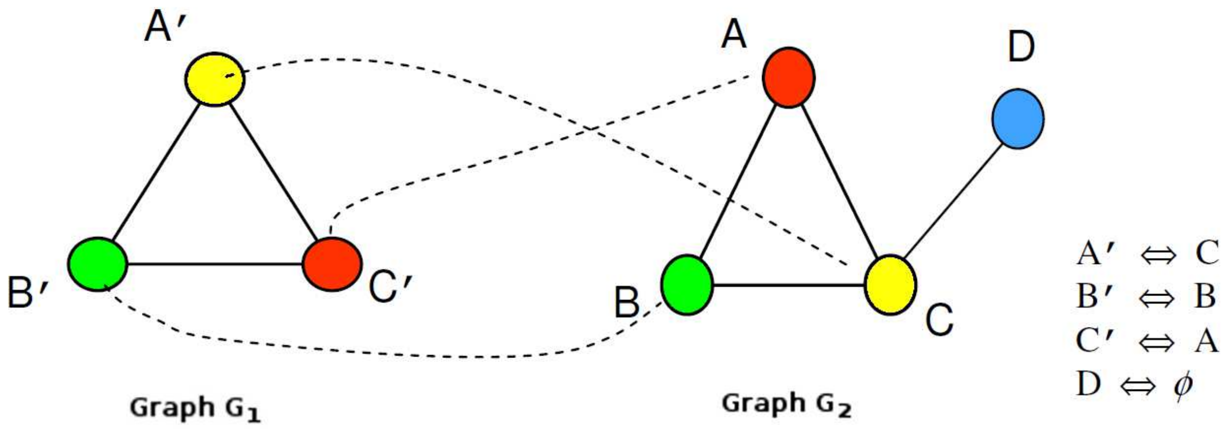
- Graph matching and graph isomorphism
- Graph edit distance
- Graph embedding

❖ Graph matching and graph isomorphism

[Messmer, 1995] [Sonbaty and Ismail, 1998]

❖ Graph edit distance

❖ Graph embedding



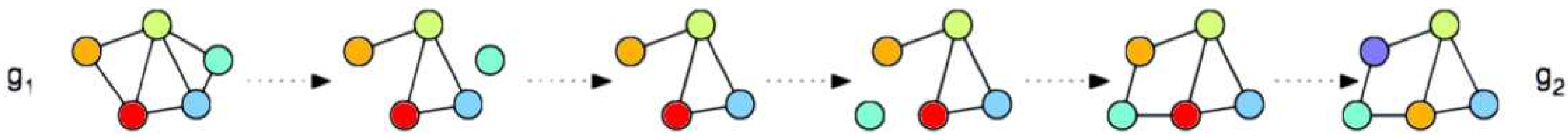
- **Graph matching and graph isomorphism**

  [Messmer, 1995] [Sonbaty and Ismail, 1998]

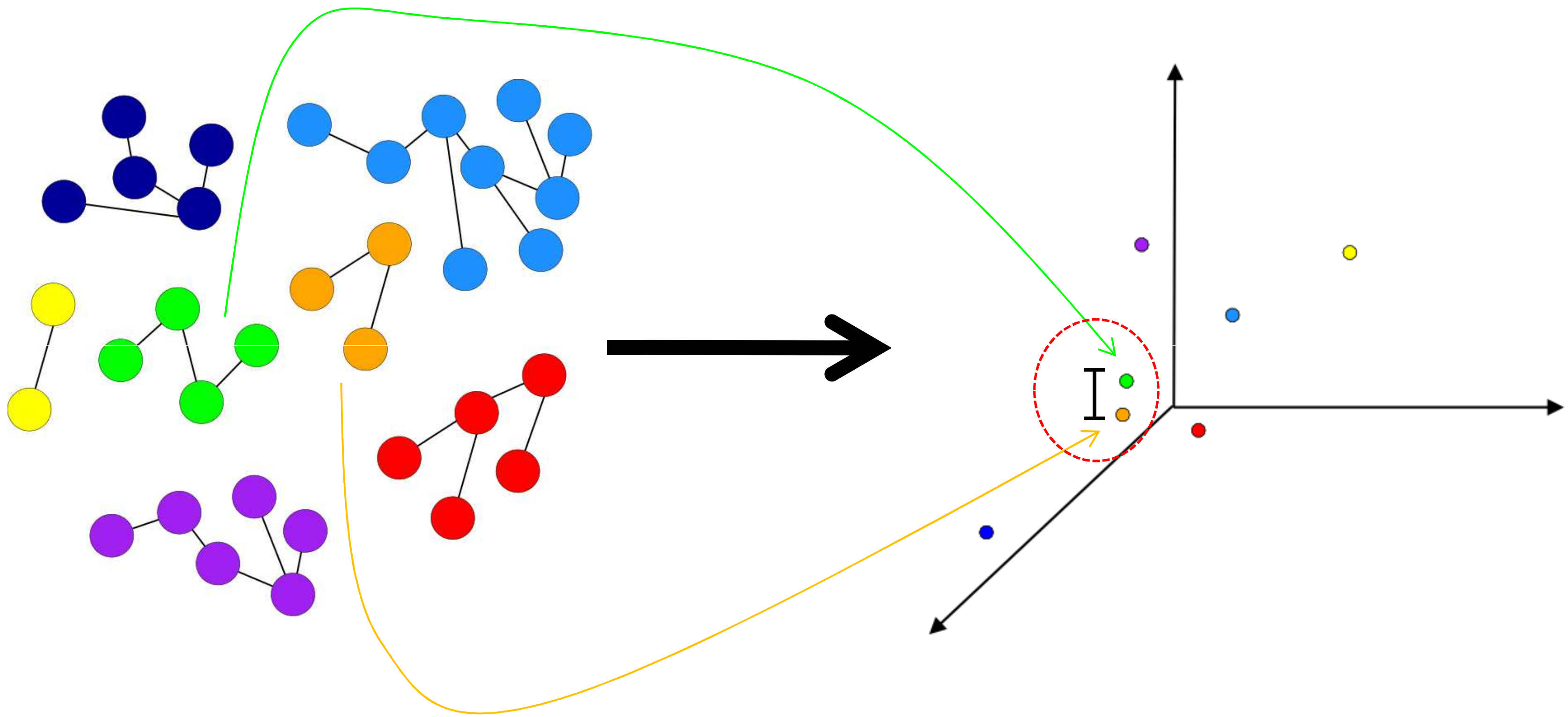
- **Graph edit distance**

  [Bunke and Shearer, 1998] [Neuhaus and Bunke, 2006]

- **Graph embedding**



Graph embedding (GEM)



Graph embedding (GEM)

Structutal PR

Expressive,
convenient,
powerful but
computationally expensive
representations

Graph embedding

Statistical PR

Mathematically sound,
mature,
less expensive and
computationally efficient
models

Explicit GEM

- embeds each input graph into a numeric feature vector
- provides more useful methods of GEM for PR
- can be employed in a standard dot product for defining an implicit graph embedding function

Implicit GEM

- computes scalar product of two graphs in an implicitly existing vector space, by using graph kernels
- does not permit all the operations that could be defined on vector spaces

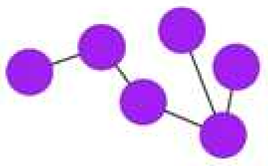
- Graph probing based methods
- Spectral based graph embedding
- Dissimilarity based graph embedding

❖ Graph probing based methods

[Wiener, 1947] [Papadopoulos et al., 1999] [Gibert et al., 2011] [Sidere, 2012]

❖ Spectral based graph embedding

❖ Dissimilarity based graph embedding



→

number of nodes = 6
number of edges = 5
etc.

v = 6,5, …

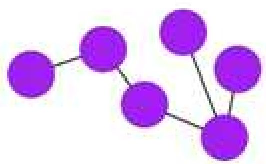
❖ Graph probing based methods

[Wiener, 1947] [Papadopoulos et al., 1999] [Gibert et al., 2011] [Sidere, 2012]

❖ Spectral based graph embedding

[Harchaoui, 2007] [Luo et al., 2003] [Robleskelly and Hancock, 2007]

❖ Dissimilarity based graph embedding



	1				
1		1			
	1		1		
		1		1	1
			1		
			1		

Spectral graph theory employing the adjacency and Laplacien matrices

Eigen values and Eigen vectors

PCA, ICA, MDS

❖ Graph probing based methods

[Wiener, 1947] [Papadopoulos et al., 1999] [Gibert et al., 2011] [Sidere, 2012]

❖ Spectral based graph embedding

[Harchaoui, 2007] [Luo et al., 2003] [Robleskelly and Hancock, 2007]

❖ Dissimilarity based graph embedding

[Pekalska et al., 2005] [Ferrer et al., 2008] [Riesen, 2010] [Bunke et al., 2011]

g

Prototype graphs
P1
P2
P3
…

v = d(g, P1), d(g, P2), …

Limitations of existing methods

- Not many methods for both directed and undirected attributed graphs
- No method explicitly addresses noise sensitivity of graphs
- Expensive deployment to other application domains
- Time complexity
- Loss of topological information
- Loss of matching between nodes
- No graph embedding based solution to answer high level semantic problems for graphs

- Introduction
- **Fuzzy Multilevel Graph Embedding (FMGE)**
  - Method
  - Experimental evaluation
  - Application to symbol recognition
  - Discussion
- Automatic indexing of graph repositories for graph retrieval and subgraph spotting
- Conclusions and future research challenges

- Fuzzy Multilevel Graph Embedding (FMGE)
- Graph probing based explicit graph embedding method

$$\phi : G \longrightarrow \mathbb{R}^n$$
$$AG \longmapsto \phi(AG) = (f_1, f_2, ..., f_n)$$

- Multilevel analysis of graph

Graph Level Information **[macro details]**	Structural Level Information **[intermediate details]**	Elementary Level Information **[micro details]**
✓ Graph order ✓ Graph size	✓ Node degree ✓ Homogeneity of subgraphs in graph	✓ Node attributes ✓ Edge attributes

- Numeric feature vector embeds a graph, encoding:
  - ✓ Numeric information by <u>fuzzy</u> histograms
  - ✓ Symbolic information by crisp histograms

FMGE

$AG_1 = (V_1, E_1, \mu^{V1}, \mu^{E1})$
$AG_2 = (V_2, E_2, \mu^{V2}, \mu^{E2})$
$\vdots$
$AG_e = (V_e, E_e, \mu^{Ve}, \mu^{Ee})$
$\vdots$
$AG_m = (V_m, E_m, \mu^{Vm}, \mu^{Em})$

$\rightarrow$ **Fuzzy Multilevel Graph Embedding (FMGE)** $\rightarrow$

$FSMFV_1$
$FSMFV_2$
$\vdots$
$FSMFV_e$
$\vdots$
$FSMFV_m$

- **Input :** Collection of attributed graphs
- **Output :** Equal-size numeric feature vector for each input graph

Fuzzy Structural Multilevel Feature Vector

Graph Level Information **[macro details]**	Structural Level Information **[intermediate details]**	Elementary Level Information **[micro details]**

Graph order	Graph size	…

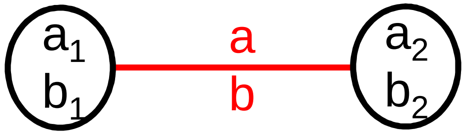
Fuzzy histogram of node degrees	Fuzzy histograms of numeric resemblance attributes	Crisp histograms of symbolic resemblance attributes	…

Fuzzy histograms of numeric node attributes	Crisp histograms of symbolic node attributes	Fuzzy histograms of numeric edge attributes	Crisp histograms of symbolic edge attributes

Homogeneity of subgraphs in a graph

- Node-resemblance for an edge
- Edge-resemblance for a node

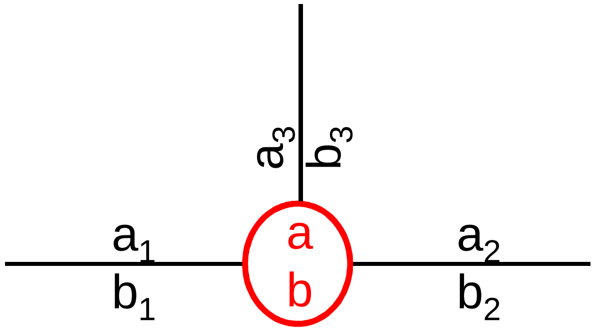
- Node-resemblance for an edge
- Edge-resemblance for a node



$$numeric\ resemblance = \frac{\min(|a_1|, |a_2|)}{\max(|a_1|, |a_2|)}$$

$$symbolic\ resemblance = \begin{cases} 1 & if\ b_1 = b_2 \\ 0 & otherwise \end{cases}$$

- Node-resemblance for an edge
- Edge-resemblance for a node



$$numeric\ resemblance = \frac{\min(|a_1|, |a_2|)}{\max(|a_1|, |a_2|)}$$

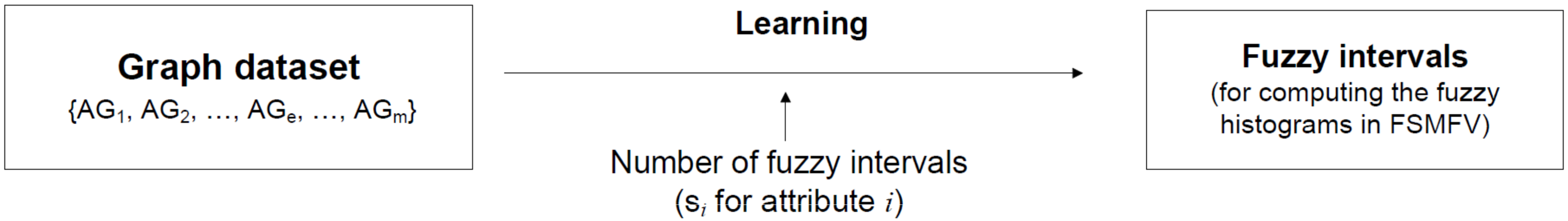
$$symbolic\ resemblance = \begin{cases} 1 & if\ b_1 = b_2 \\ 0 & otherwise \end{cases}$$

$$\frac{(b_1, b_2) + (b_1, b_3) + (b_2, b_3)}{3}$$

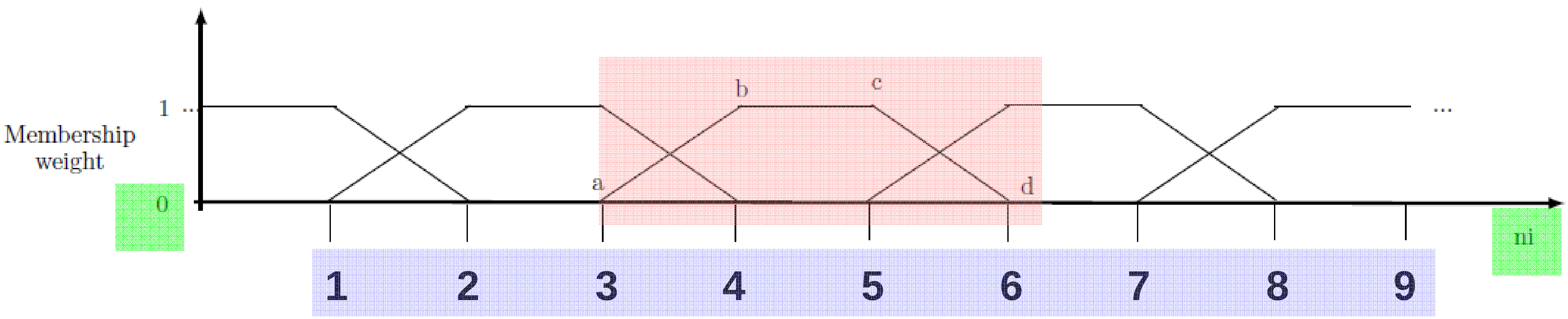
$$\frac{(a_1, a_2) + (a_1, a_3) + (a_2, a_3)}{3}$$

- Unsupervised learning phase
- Graph embedding phase

Unsupervised learning phase of FMGE



List of values of attribute$_i$ in input collection of graphs → Discretization → Crisp intervals → Crisp to fuzzy intervals → 's$_i$' fuzzy intervals for attribute$_i$



- First fuzzy interval (-∞, -∞, …, …)
- Last fuzzy interval (…, …, ∞, ∞)

Graph
$AG_e = (V_e, E_e, \mu^{Ve}, \mu^{Ee})$

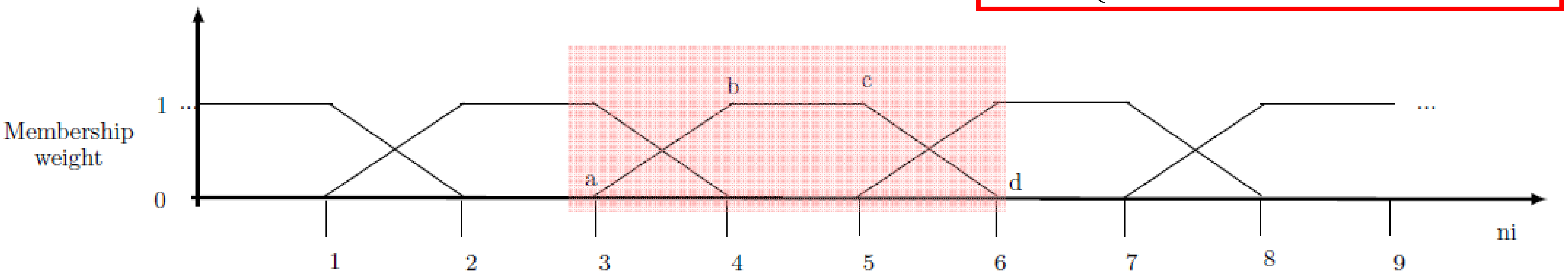
Embedding →

Fuzzy overlapping trapezoidal intervals
(s_i for attribute i)

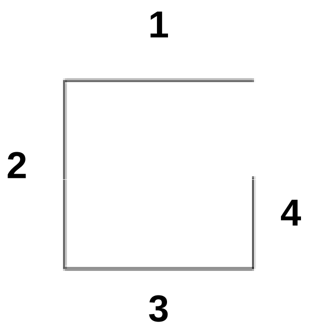
Fuzzy Structural Multilevel Feature Vector
($FSMFV_e$)

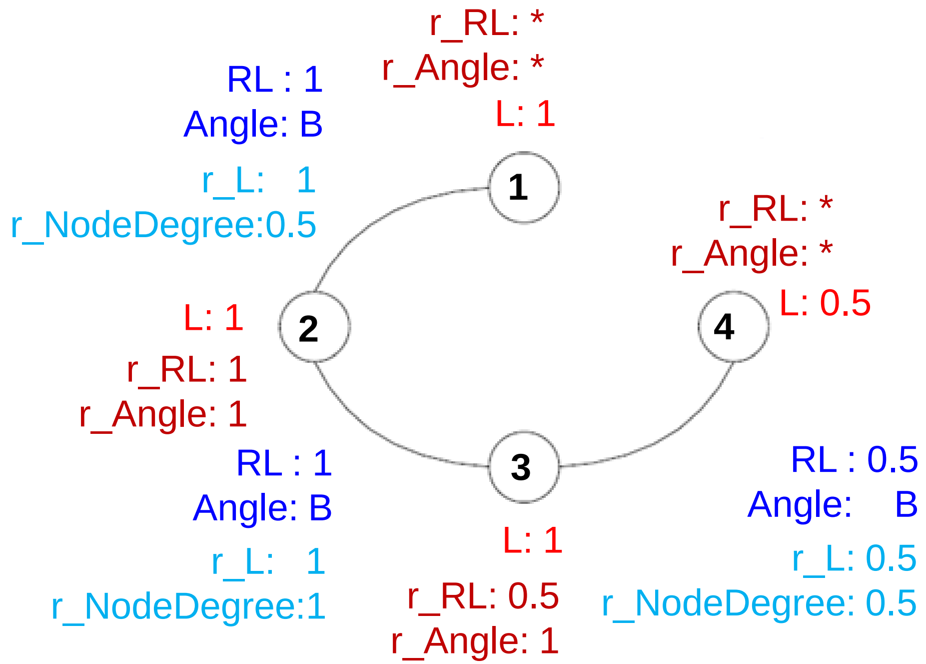
- Numeric information embedded by fuzzy histograms
- Symbolic information embedded by crisp histograms

$$\alpha(x)=\begin{cases}(x-a)/(b-a) & \text{if } a\le x<b\\ 1 & \text{if } b\le x\le c\\ (x-d)/(c-d) & \text{if } c<x\le d\\ 0 & \text{otherwise}\end{cases}$$

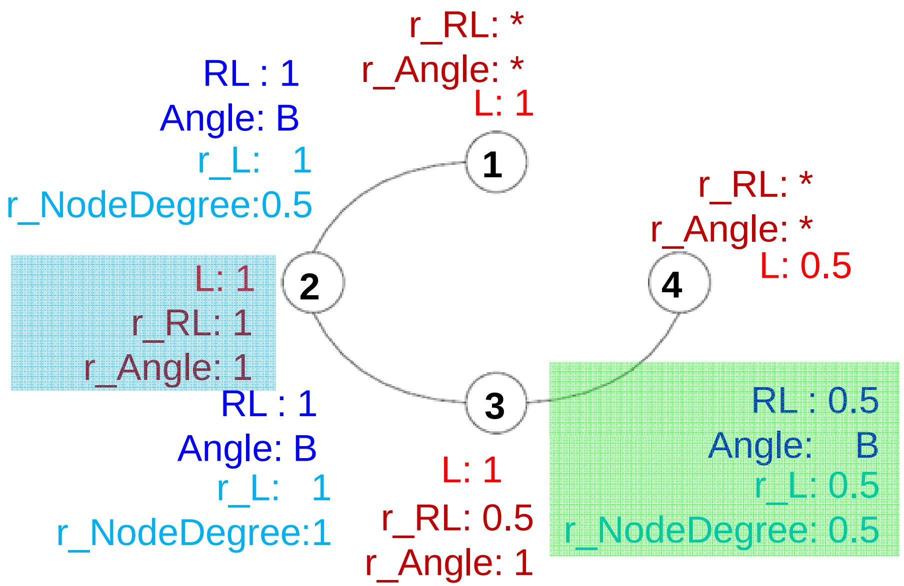


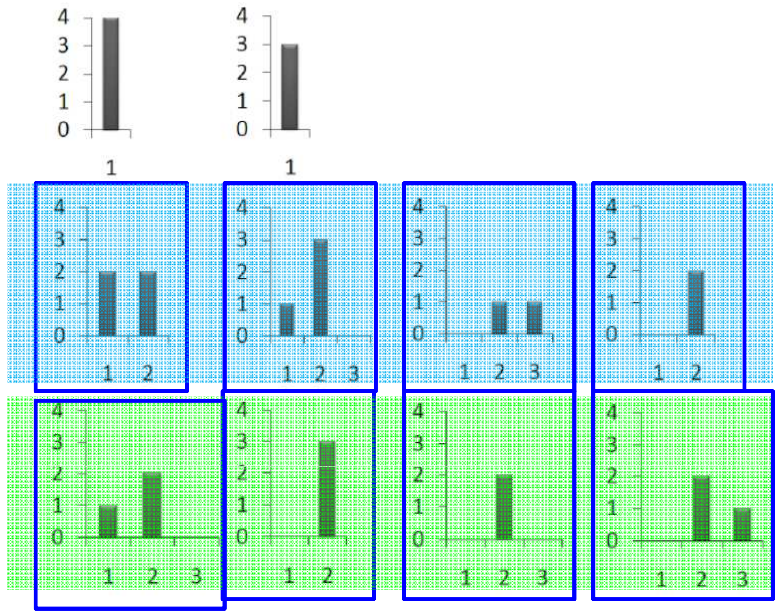
Example - FMGE





Example - FMGE





FSMFV: 4,3,2,2,1,3,0,0,1,1,0,2,1,2,0,0,3,0,2,0,0,2,1

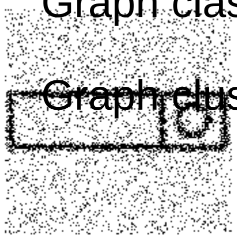
- Node degree: [-∞,-∞,1,2] and [1,2,∞,∞]
- Attributes {*L,RL*}: [-∞,-∞,0.5,1], [0.5,1,1.5,2] and [1.5,2, ∞,∞]
- r_Angle: [-∞,-∞,0,1] and [0,1, ∞,∞]
- Resemblance attributes: [-∞,-∞,0.25,0.5], [0.25,0.5,0.75,1.0] and [0.75,1.0, ∞,∞,]
- The symbolic edge attribute Angle has two possible labels

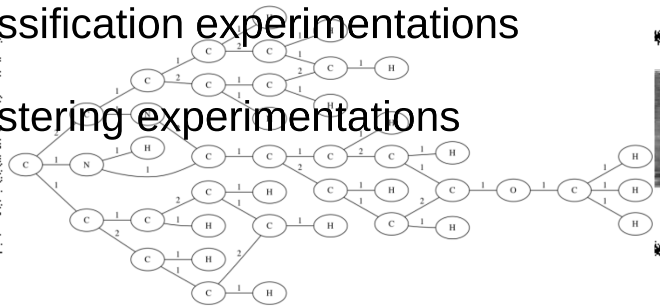
Experimental evaluation of FMGE

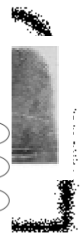
Dataset	Size			Classes	Avg		Max		Attributes[a]	
	Train	Valid	Test		$\|V\|$	$\|E\|$	$\|V\|$	$\|E\|$	V	E
Letter LOW	750	750	750	15	4.7	3.1	8	6	2;0	0;0
Letter MED	750	750	750	15	4.7	3.2	9	7	2;0	0;0
Letter HIGH	750	750	750	15	4.7	4.5	9	9	2;0	0;0
GREC	836	836	1628	22	11.5	12.2	25	30	2;1	1;1
Fingerprint	500	300	2000	4	5.4	4.4	26	25	2;0	1;0
Mutagenicity	500	500	1500	2	30.3	30.8	417	112	0;1	1;0

[a] Number of attributes is given as a pair "numeric;symbolic".

- IAM graph database
  - ✓ Graph classification experimentations
  - ✓ Graph clustering experimentations







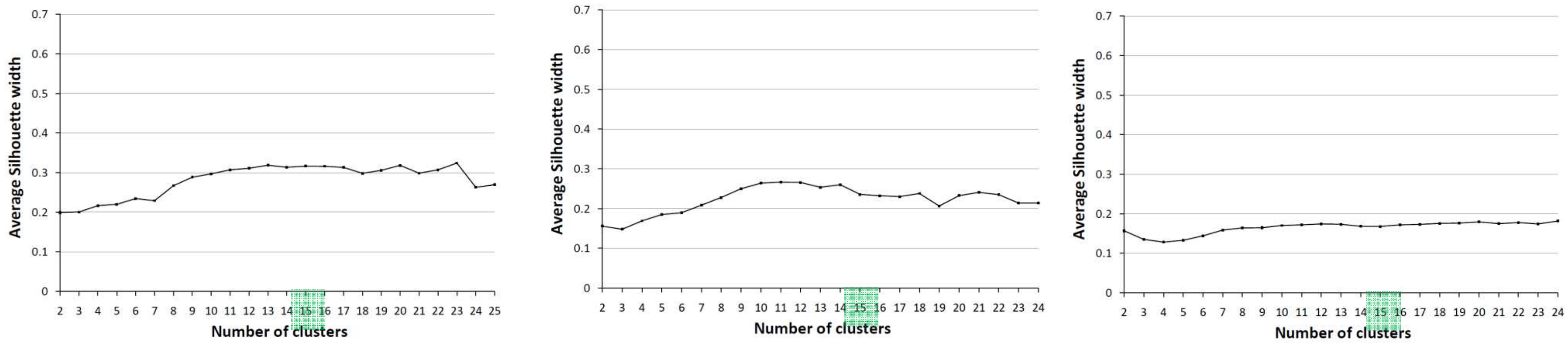
Dataset	Graph edit distance based reference system [k-NN classifier]	Dissimilarity based embedding Bunke et al. [Bunke and Riesen, 2011b] [SVM classifier]	FMGE resemblance:AVG [1-NN classifier]	FMGE resemblance:STD [1-NN classifier]
Letter LOW	99.3	99.3	97.1	97.1
Letter MED	94.4	94.9	75.7	75.7
Letter HIGH	89.1	92.9	60.5	60.5
GREC	82.2	92.4	**97.5**	**97.5**
Fingerprint	79.1		74.9	73.5
Mutagenicity	66.9		**68.6**	**68.6**

- Supervised machine learning framework for experimentation, employing the training, validation and test sets
- 1-NN classifier with Euclidean distance.
- Equal-spaced crisp discretization and the number of fuzzy intervals empirically selected on validation dataset

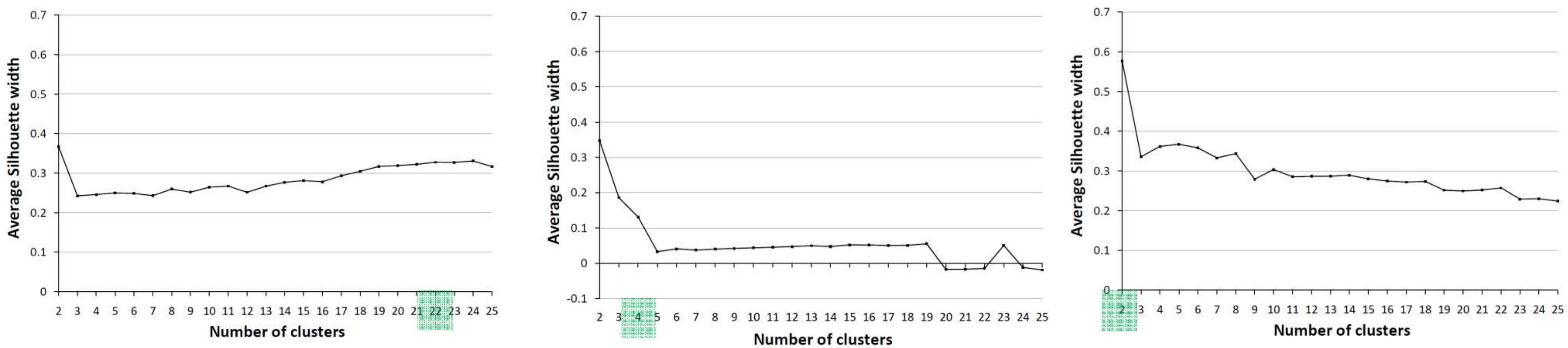
Dataset	FMGE feature vector space
	correctly clustered graphs (%)
Letter LOW	89
Letter MED	60
Letter HIGH	41
GREC	82
Fingerprint	57
Mutagenicity	82

- Merged training, validation and test sets
- K-means clustering with random non-deterministic initialization
- The measure of quality of K-means clustering w.r.t. the ground truth : ratio of correctly clustered graphs to the graphs in the dataset
- Equal-frequency crisp discretization for automatically selecting the best number of fuzzy intervals

Graph clustering experimentations



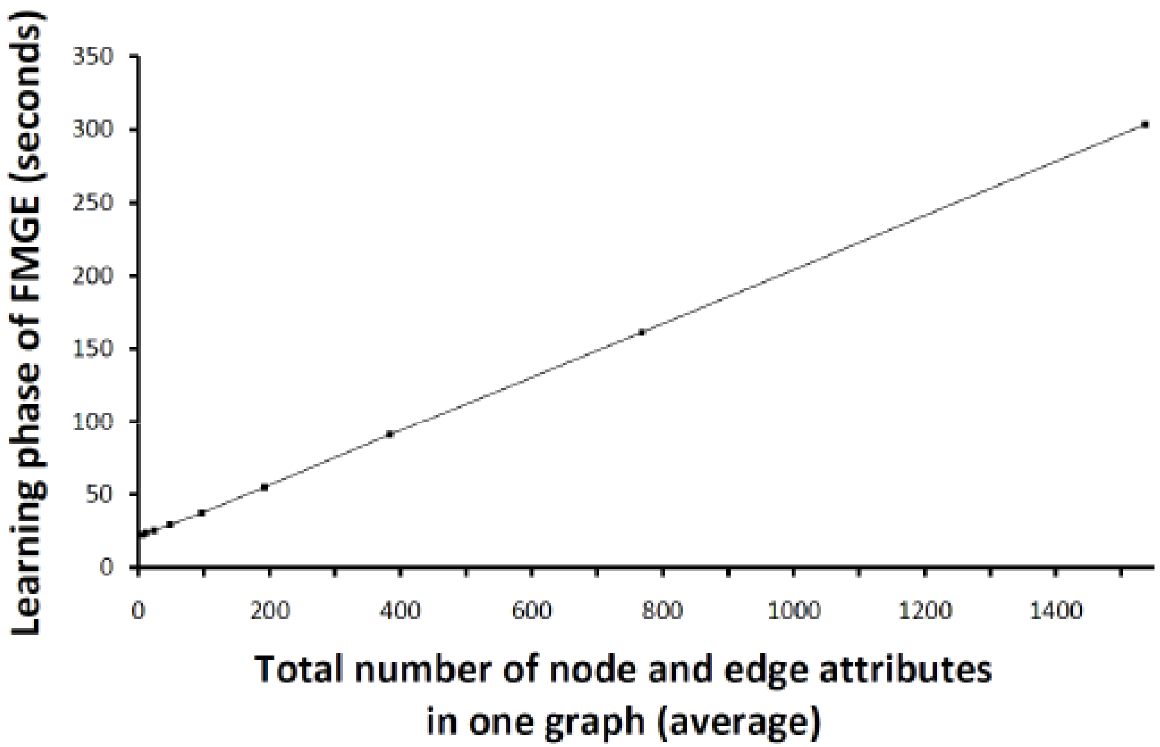
Letter-LOW, Letter-MED and Letter-HIGH



GREC, Fingerprint and Mutagenicity

- The average Silhouette width ranges between [-1, 1]. The closer it is to 1, the better the is the clustering quality.

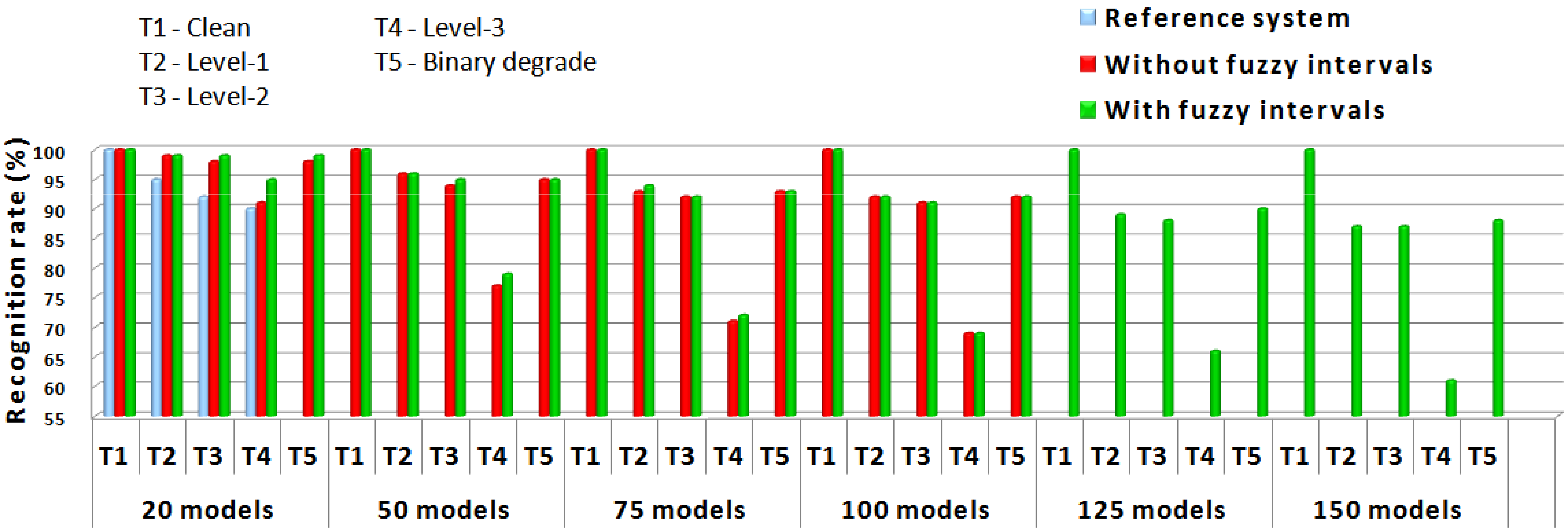
Time complexity of FMGE

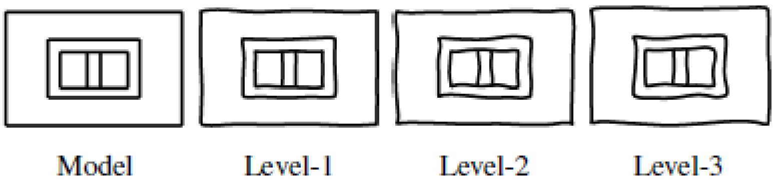


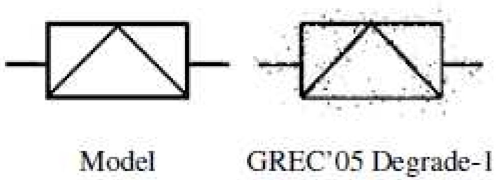
- Unsupervised learning phase is performed off-line and is linear to:
  - ✓ Number of node and edge attributes
  - ✓ Size of graphs
- Graph embedding phase is performed on-line

Application to symbol recognition

- 2D linear model symbols from GREC databases
- Learning on clean symbols and testing against noisy and deformed symbols

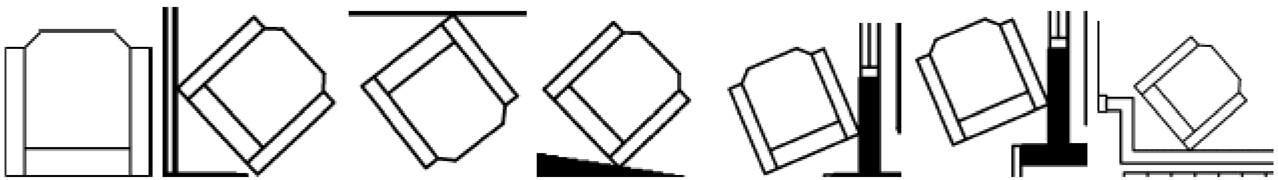






- SESYD dataset
- Learning on clean symbols and testing against noisy symbols

	Noise	Model symbol (classes)	Query symbol (each class)	Recognition rate (match with topmost result)	Recognition rate (a match in top-3 results)
Floor plans	Level-1	16	100	84%	95 %
	Level-2	16	100	79%	90 %
	Level-3	16	100	76%	87 %
Average recognition rate				80%	91%
Electronic diagrams	Level-1	21	100	69%	89%
	Level-2	21	100	66%	88%
	Level-3	21	100	61%	85%
Average recognition rate				65%	87%



Model = Level-1 = = Level-2 = = Level-3 =

- Not many methods for both directed and undirected attributed graphs
  - ✓ FMGE: Directed and undirected graphs with many numeric as well as symbolic attributes on both nodes and edges
- No method explicitly addresses noise sensitivity of graphs
  - ✓ FMGE: Fuzzy overlapping intervals
- Expensive deployment to other application domains
  - ✓ FMGE: Unsupervised learning abilities

- Time complexity
  - ✓ FMGE: Linear to number of attributes
    Linear to size of graphs
    Graph embedding performed on-line
- Loss of topological information
  - ✓ FMGE: Multilevel information (global, topological and elementary)
    Homogeneity of subgraphs in graph

Summary and discussion - FMGE

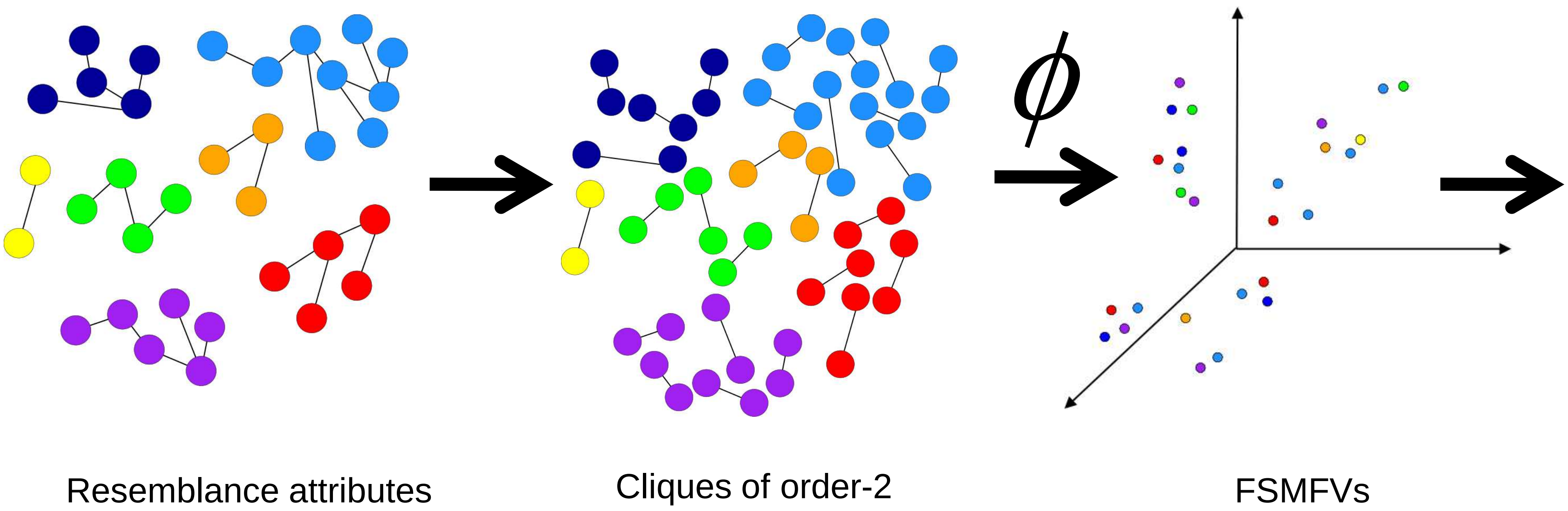
- Loss of matching between nodes
- No graph embedding based solution to answer high level semantic problems for graphs

- Introduction
- Fuzzy Multilevel Graph Embedding (FMGE)
- **Automatic indexing of graph repositories for graph retrieval and subgraph spotting**
  - Method
  - Experimental evaluation - application to content spotting in graphic document image repositories
  - Discussion
- Conclusions and future research challenges

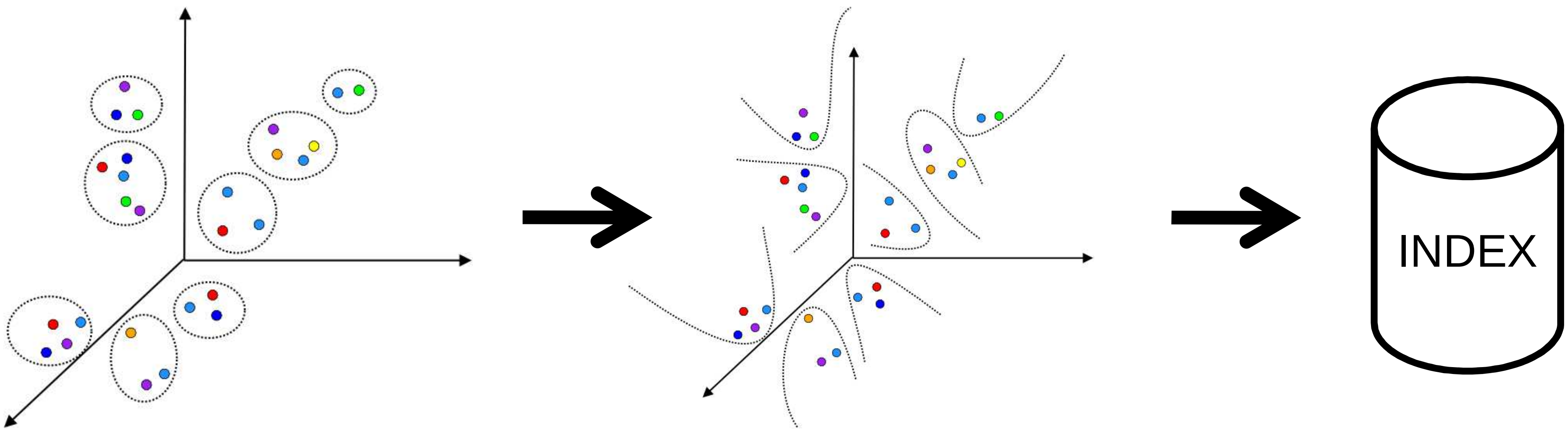
- Bag of words inspired model for graphs
- Index the graph repository by elementary subgraphs
- Explicit GEM for exploiting computational strengths of state of the art machine learning, classification and clustering tools

- Unsupervised indexing phase
- Graph retrieval and subgraph spotting phase

- **Unsupervised indexing phase**
- Graph retrieval and subgraph spotting phase



- **Unsupervised indexing phase**
- Graph retrieval and subgraph spotting phase

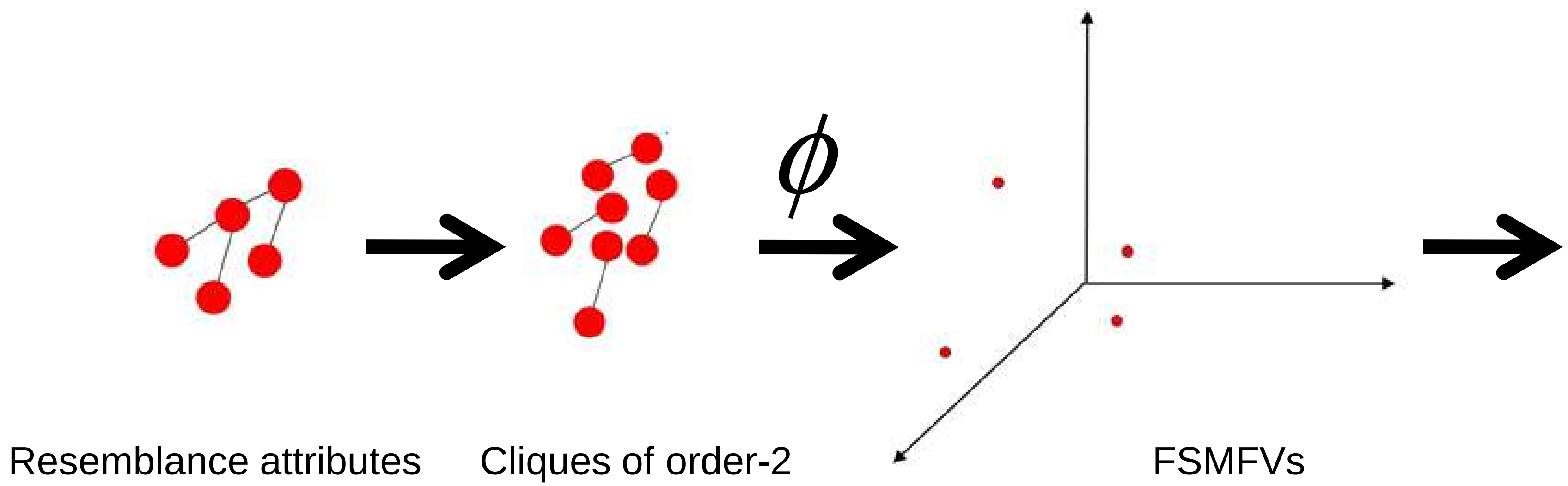


FSMFV clusters using an hierarchical clustering technique

Classifier

- Unsupervised indexing phase

- **Graph retrieval and subgraph spotting phase**



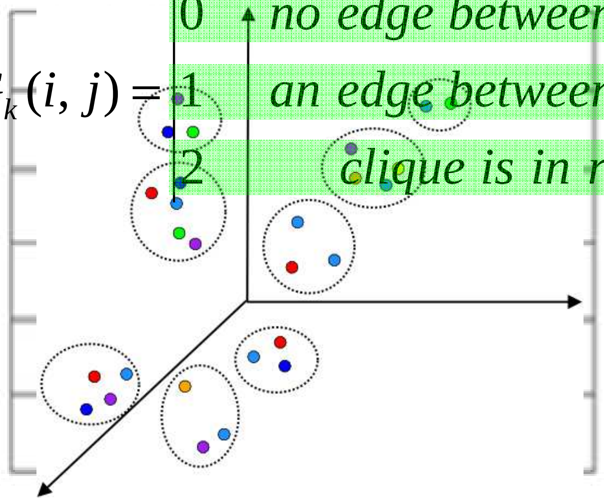
- Unsupervised indexing phase

- **Graph retrieval and subgraph spotting phase**

Adjacency matrix of a result

$$AG_k(i,j) = \begin{cases} 0 & \textit{no edge between } i \textit{ and } j \\ 1 & \textit{an edge between } i \textit{ and } j \\ 2 & \textit{clique is in result} \end{cases}$$

where,



Classify

clique for which score is computed

clique having the same cluster label as one of cliques in query graph

clique not having the same cluster label as one of the cliques in the query graph

no edge in graph

INDEX

Set of result graphs

$$score = \sum_{z=0}^{2} \left(z \times \frac{|z|}{w} \right)$$

z is a value in adjacency matrix (either 0, 1, 2)

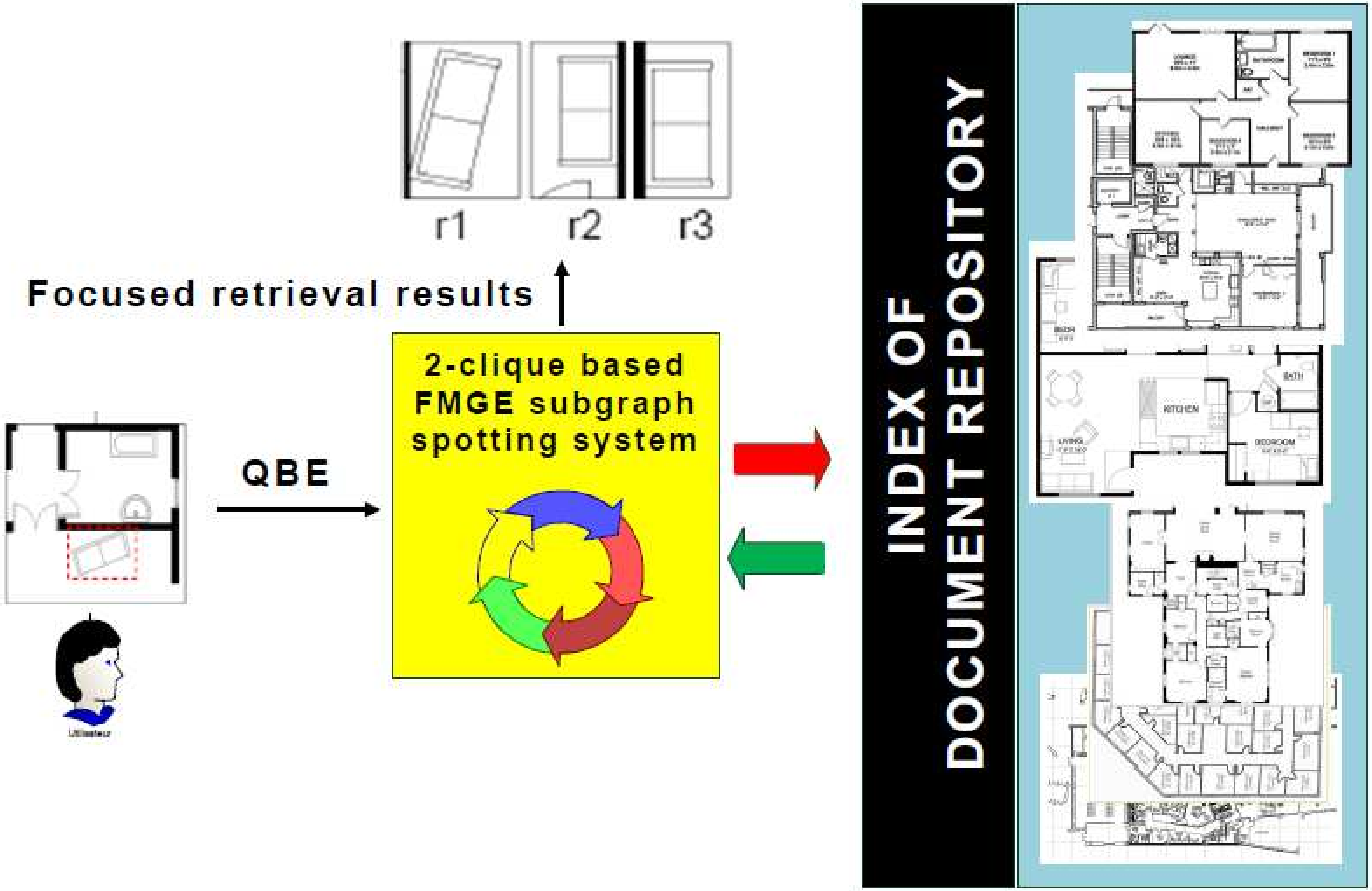
|z| is frequency of value z in neighborhood

and

w is number of connected neighbors looked-up

score = (0 * 4 / 9) + (1 * 2 / 9) + (2 * 3 / 9) = 0.89

Content spotting in document images

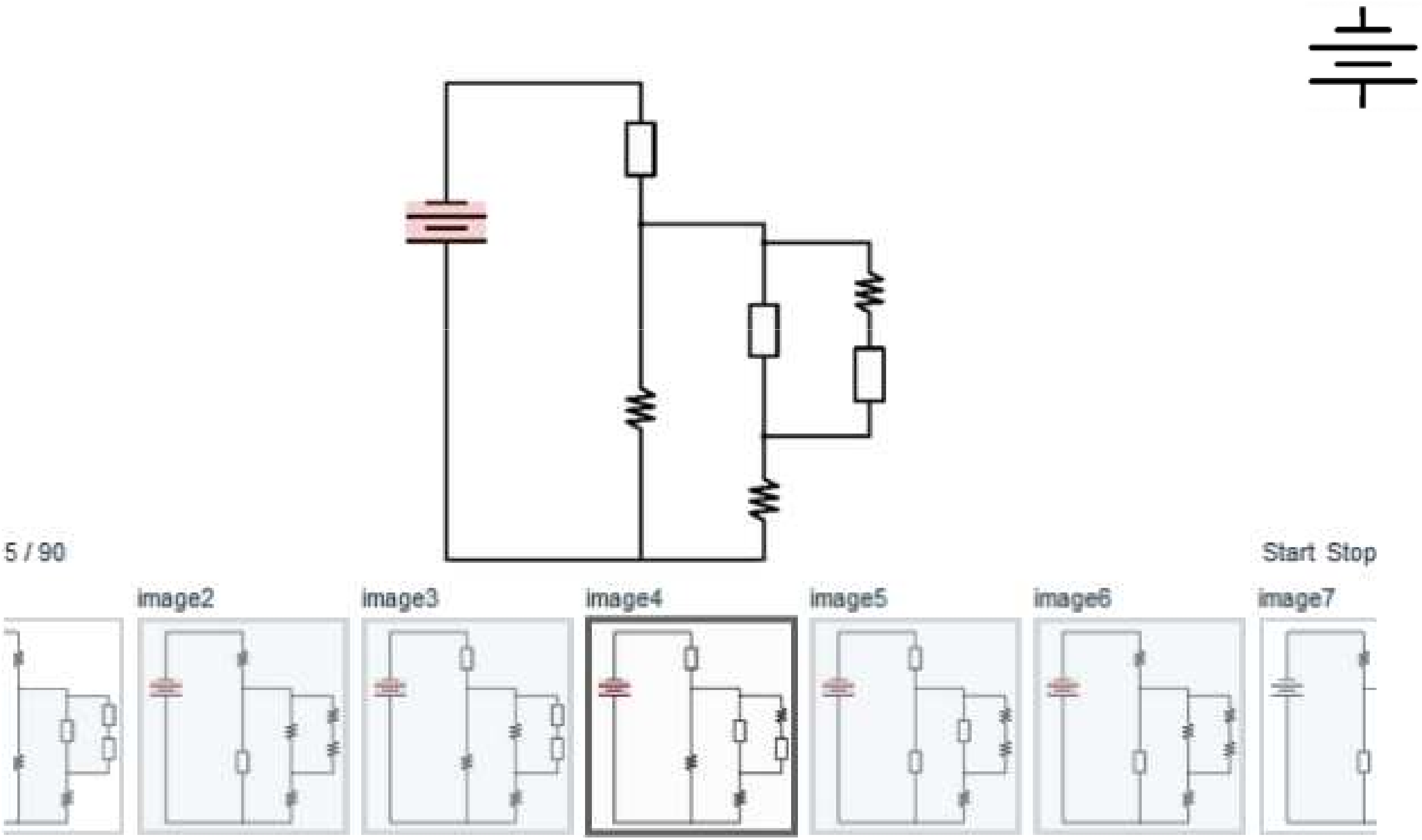


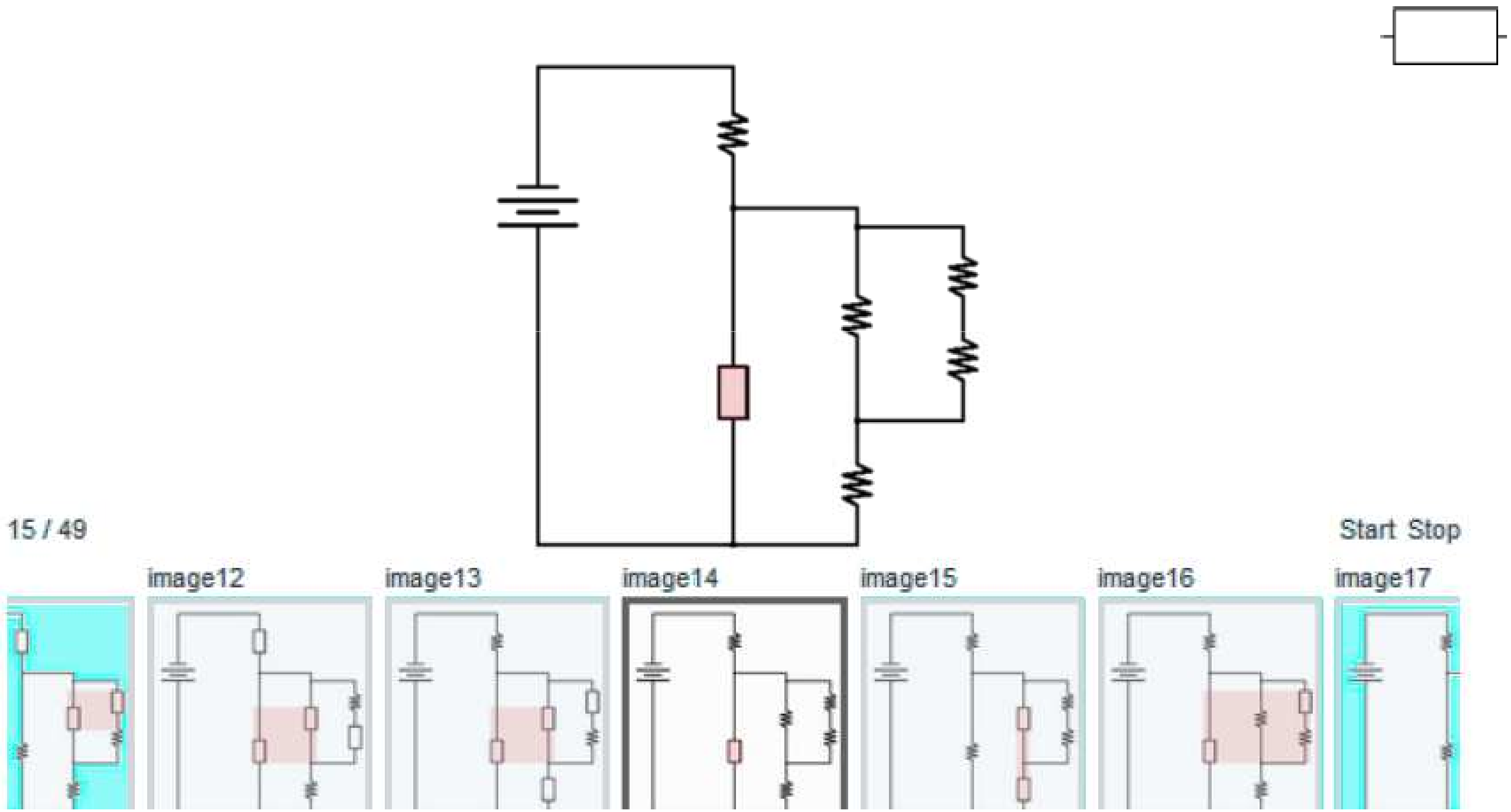
- SESYD dataset
- Corresponding graph dataset is made publically available

  http://www.rfai.li.univ-tours.fr/PagesPerso/mmluqman/public/SESYD_graphs.zip

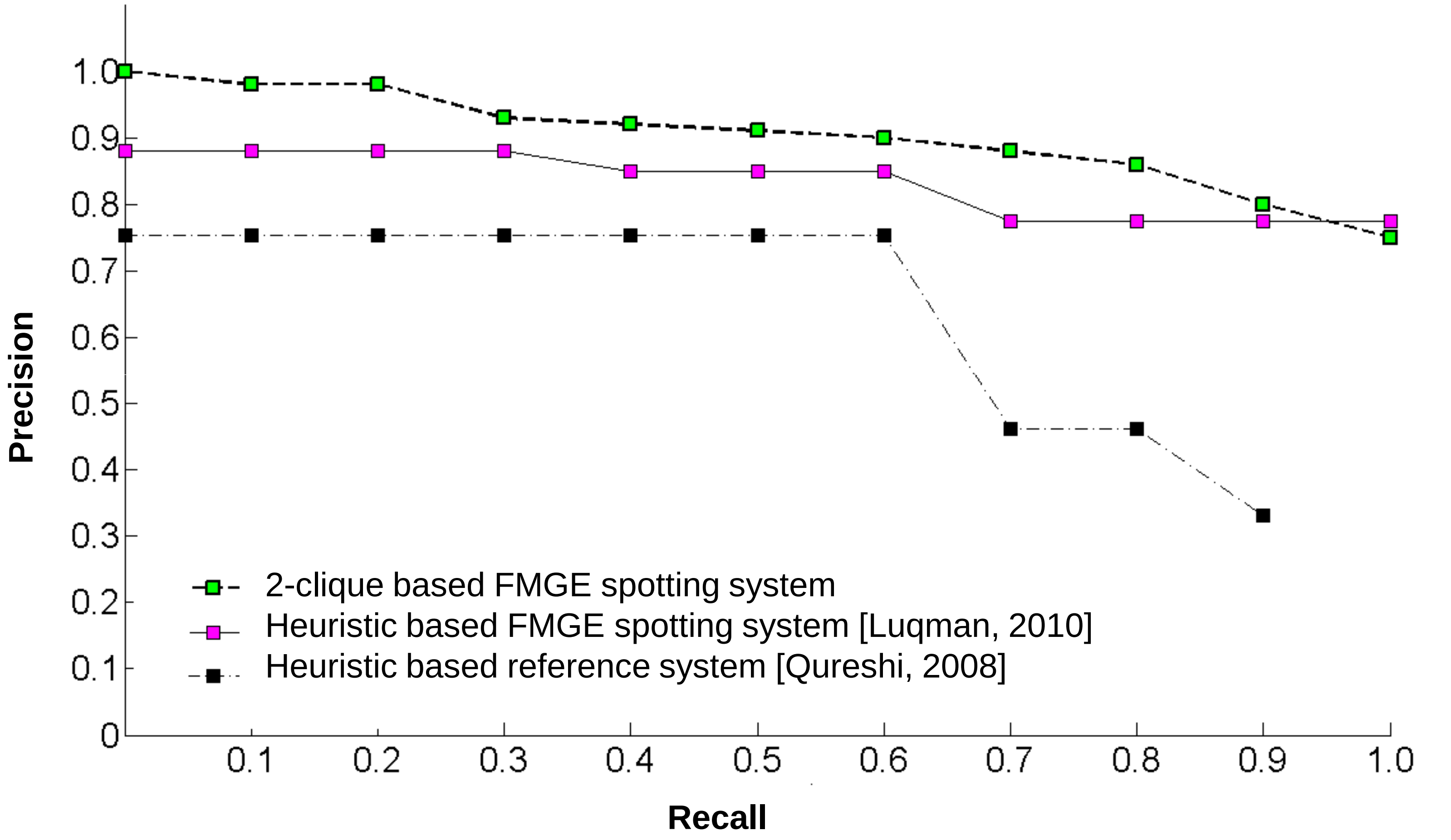
	Image		Attributed graph	
Electronic diagrams	Backgrounds	8	Avg. order	212
	Models	21	Avg. size	363
	Symbols	9600	Node attribs.	4
			Edge attribs.	2
	Documents	800	Graphs	800
	Queries	1000	Graphs	1000
Architectural floor plans	Backgrounds	2	Avg. order	359
	Models	16	Avg. size	733
	Symbols	4216	Node attribs.	4
			Edge attribs.	2
	Documents	200	Graphs	200
	Queries	1000	Graphs	1000

Experimental evaluation

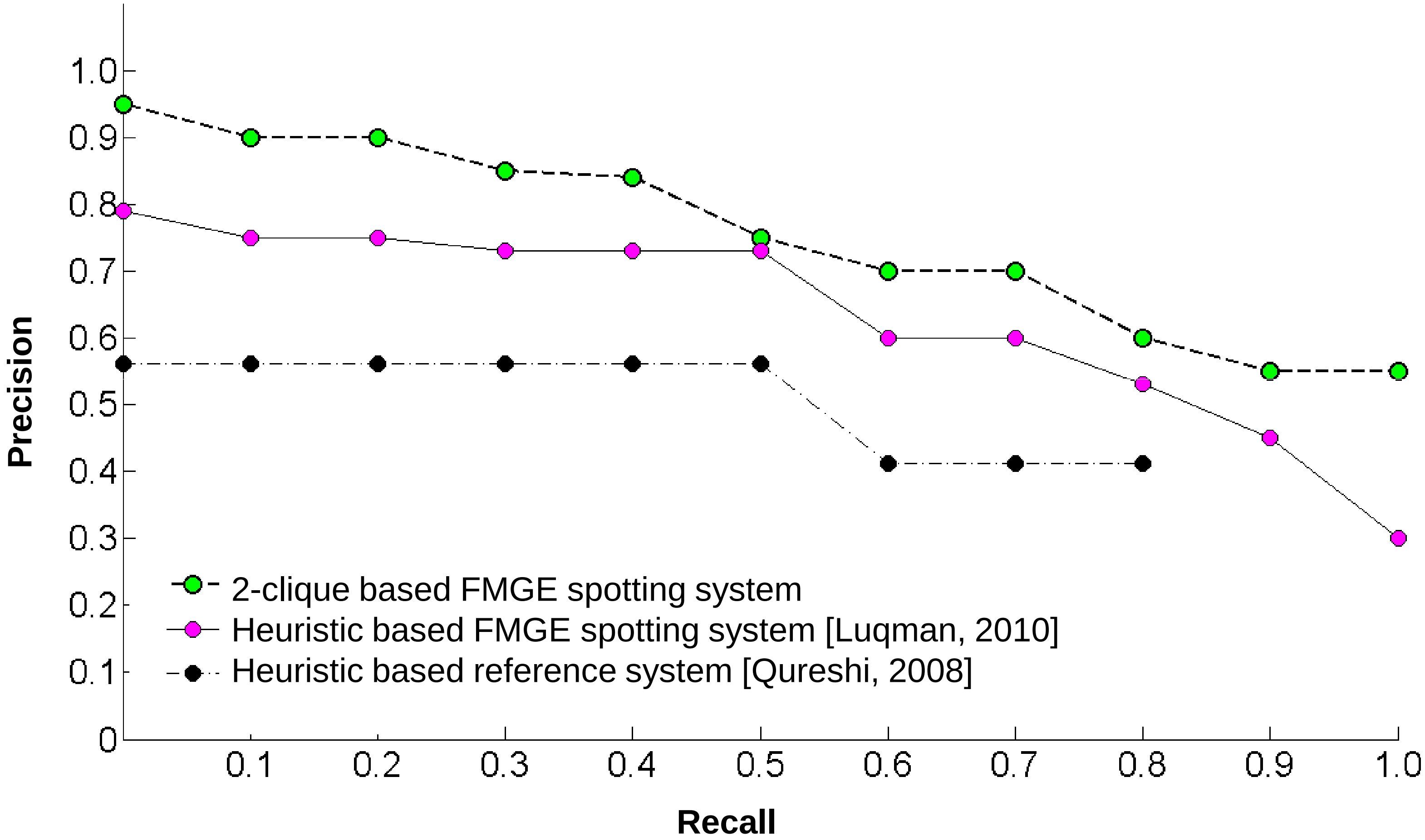



15 / 49
Start Stop
image12
image13
image14
image15
image16
image17

Experimental evaluation



Electronic diagrams: (517K 2-node subgraphs) (455 classes) (~17h)



Architectural diagrams: (306K 2-node subgraphs) (211 classes)

Discussion – subgraph spotting

- Loss of matching between nodes
  - ✓ Score function is a first step forward
- No graph embedding based solution to answer high level semantic problems for graphs
  - ✓ FMGE based framework for automatic indexing of graph repositories

- Introduction
- Fuzzy Multilevel Graph Embedding (FMGE)
- Automatic indexing of graph repositories for graph retrieval and subgraph spotting
- **Conclusions and future research challenges**

- Last two decade's research on structural pattern recognition can access state of the art machine learning tools
- An impossible operation in original graph space turns into a realizable operation with an acceptable accuracy
- Application to domains where the use of graphs is mandatory for representing rich structural and topological information and a computational efficient solution is required
- Feature vector not capable of preserving the matching between nodes of a pair of graphs

- Unsupervised and automatic indexing of graph repositories
- Domain independent framework
- Incorporating learning abilities to structural representations
- Ease of query by example (QBE)
- Granularity of focused retrieval

Ongoing and short term

- Dimensionality reduction
- Feature selection
- More topological information

Medium term

- Detection of outliers for cleaning learning set
- Multi-resolution index using cliques of higher order (≥3)

Future research challenges

❖ **Long term**

- Surjective mapping of nodes of two graphs

Journal paper Pattern Recognition (under review, submitted December 2011)	**1**
Book chapter Bayesian Network by InTech publisher	**1**
International conference contributions ICDAR 2011, ICPR 2010, ICDAR 2009	**3**
Selected papers for post-workshop LNCS publication ICPR 2010 contests, GREC 2009	**2**
International workshop contributions GREC 2011, GREC 2009	**2**
Francophone conference contributions CIFED 2012, CIFED 2010	**2**

Thank you for your attention.

Cotutelle PhD thesis

Fuzzy Multilevel Graph Embedding for Recognition, Indexing and Retrieval of Graphic Document Images

presented by

Muhammad Muzzamil LUQMAN

mluqman@{univ-tours.fr, cvc.uab.es}

Friday, 2nd of March 2012

Directors of thesis

Dr. Jean-Yves RAMEL
Professor
University of Tours, France

Dr. Josep LLADOS
Professor
UAB, Spain

Co-supervisor

Dr. Thierry BROUARD
Assistant Professor
University of Tours, France