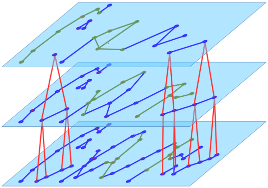
Pyramidal Stochastic Graphlet Embedding for Document Pattern Classification

Anjan Dutta, Pau Riba, Josep Lladós, Alicia Fornés

Computer Vision Center, Autonomous University of Barcelona

ICDAR, Kyoto, Japan, 13th November, 2017



Horizon 2020
European Union funding
for Research & Innovation

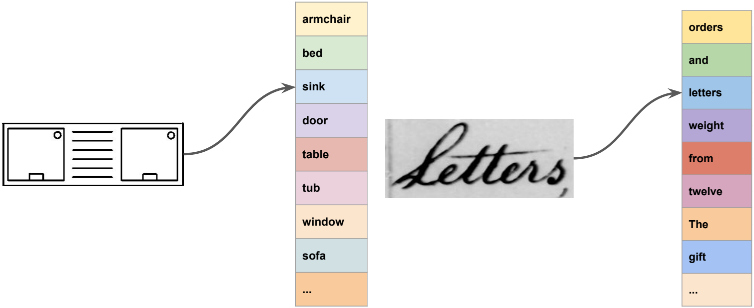
Outline

Introduction

Introduction

Document pattern classification

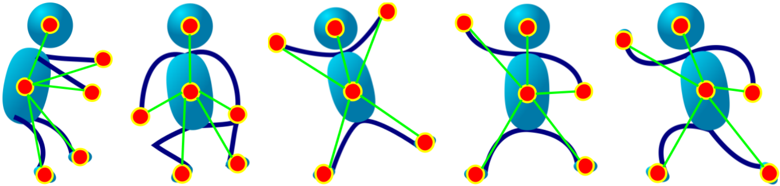
- Word and symbol classification.
- Application: document feature generation, document categorization, spam filtering etc.



Introduction

Graph based representation

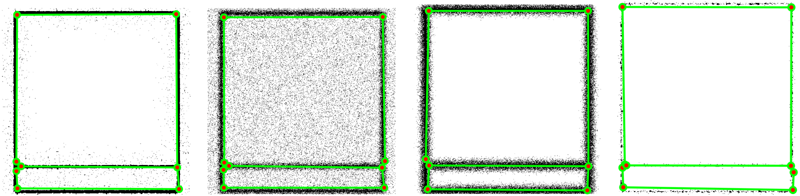
- Limitations of statistical pattern recognition.
- Advantages of structural pattern recognition.
- Graph based representation: relation between object parts.
- Invariant to rotation and affine transformation.
- Comparing graphs: graph matching, graph kernel.



Introduction

Motivation

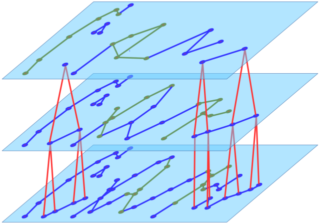
- Document part $\rightarrow$ graph $\Rightarrow$ noisy conversion
- Unstable representation.

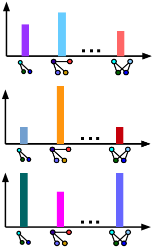


Introduction

Contribution

- Graph pyramid: multi-scale graph, tolerate noise, stable representation.
- Stochastic graphlet embedding: avoid graph matching, allows application of machine learning techniques, low to high order graphlets statistics.

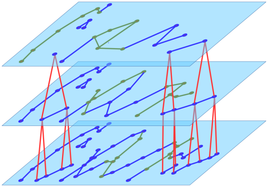




Pyramidal Graph Representation

Pyramidal Graph Representation

- Multi-scale graph, information at different resolutions.
- Higher leveled graphs contain abstract information.
- Graph pyramid construction techniques:
  1. Girvan-Newman
  2. grPartition



Girvan-Newman Algorithm

- Algorithm for graph clustering (Girvan and Newman NAS 2002).
- Basic principle:
  1. Compute edge centrality.
  2. Remove edge with highest score.
  3. Recompute all scores.
  4. Repeat 2^{nd} step.
- Results in a dendogram where each node is an independent cluster.
- Algorithm stops when the given number of clusters is reached.

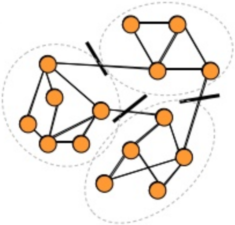
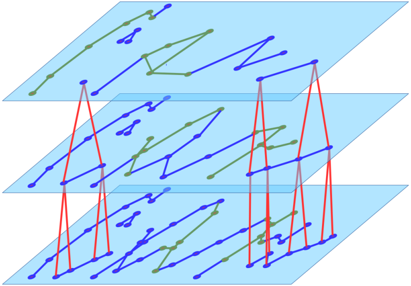


Figure credit: S. Papadopoulos, CERTH-ITI, 2011.

Pyramid Generation

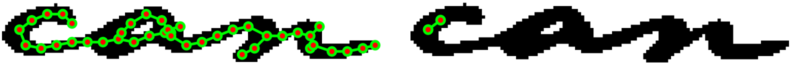
- Pyramid construction: at a higher level each cluster is represented as a node.
- Hierarchical edges: clustered nodes to their representative in the higher level.



Stochastic Graphlet Embedding

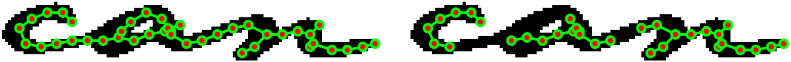
Stochastic Graphlets Sampling

- Graphlet sampling is a stochastic and recurrent procedure.
- It is controlled by two parameters M and T.
- Basic principles:
  1. Randomly select a node v from G.
  2. Add the node v to an empty graph $\mathcal{G}$.
  3. Recursively add T connected edges to $\mathcal{G}$.
  4. Restart 1$^{\text{st}}$ step M times.
- Animation: $M = 10$, $T = 6$.



Stochastic Graphlets Sampling

- A random walk process with a restart.
- Samples $M \times T$ connected graphlets, with edges varying from 1 to T.
- Hypothesis: empirical distribution of large amount of sampled graphlets will be same to actual distribution.

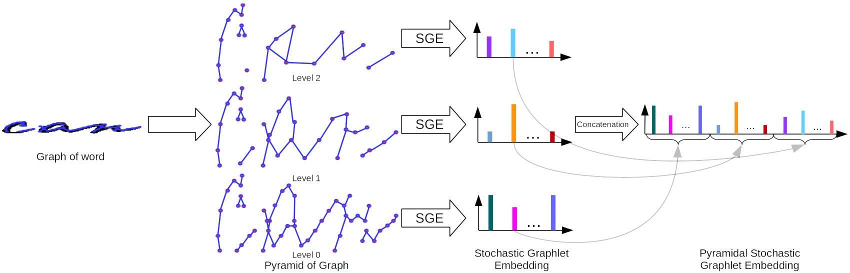


Hashed Graphlets Distribution

- Graph hash functions:
  1. Degree of nodes
  2. Betweenness centrality
  3. Core numbers
  4. Clustering coefficients
- Probability of collision (Dutta and Sahbi, ArXiv, 2017)
- Hash functions with low probability of collision: degree of nodes, betweenness centrality.
- Hash function $= \begin{cases} \text{degree of nodes,} & \text{if } t \leq 4 \\ \text{betweenness centrality,} & \text{otherwise} \end{cases}$

Pyramidal Stochastic Graphlet Embedding

Summary

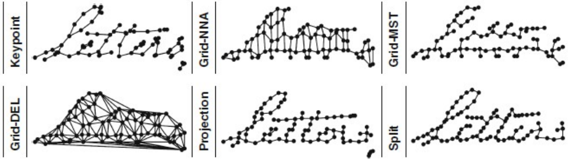


Experimental Validation

Datasets

HistoGraph

- Perfectly segmented word images from George Washington (GW) dataset.
- 30 different words and six different representations:



- Three independent subsets: training (90 words), validation (60 words) and test (143 words).
- Frequency: train and validation set (2 to 3), test set (3 to 5).

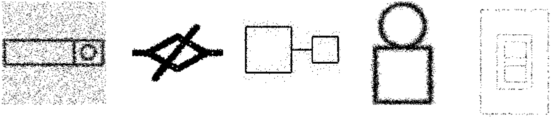
Figure credit: Stauffer *et al.* S+SSPR 2016

Results
HistoGraph

Subset	Acc. GED	Acc. SGE	Acc. PSGE	
			Level 2	Level 3
Keypoint	77.62	78.32	**80.42 (+2.10)**	78.32 (+0.00)
Grid-NNA	65.03	72.73	72.73 (+0.00)	**74.13 (+1.40)**
Grid-MST	74.13	**76.92**	75.52 (-1.40)	74.83 (-2.09)
Grid-DEL	62.94	74.83	**79.02 (+4.19)**	**79.02 (+4.19)**
Projection	**81.82**	79.02	79.72 (+0.70)	80.42 (+1.40)
Split	80.42	77.62	**80.42 (+2.80)**	77.62 (+0.00)

Datasets
GREC

- Graphs representing symbols from architectural and electronic drawings.
- 22 different classes and five different distortion levels:



- Preprocessing applied for cleaning the images and converting them to graphs.
- Three independent subsets: training and validation (286 symbols), test (528 symbols).
- Frequency: train and validation set (13), test set (24).

Figure credit: Riesen and Bunke SSPR 2008

Results

GREC

Method	Unlabelled	Labelled	
Dissimilarity Embedding (Bunke and Riesen PR 2010)	-	95.10	
Node Attribute Statistics (Gibert *et al.* PR 2012)	-	99.20	
Fuzzy Graph Embedding (Luqman *et al.* PR 2013)	-	97.30	
SGE (Dutta and Sahbi ArXiv 2017)	92.80	99.62	
		Level 2	Level 3
PSGE	93.18 (+0.38)	99.62 (+0.00)	**99.81 (+0.19)**

Conclusions and Future Work

Conclusions and Future Work

- Proposal of pyramidal stochastic graphlet embedding.
- Pyramidal representation of graph tolerates noise and distortion.
- SGE samples low to high order graphlets providing robust structural statistics.
- Consideration of hierarchical edges as a future line of work.

Thanks for your attention!

Questions?

Anjan Dutta, PhD
Marie-Curie Postdoctoral Fellow
Computer Vision Center
Autonomous University of Barcelona
Email: adutta@cvc.uab.es