Geometric-based Symbol Spotting, with Application to Symbol Retrieval in Document Image Databases

Nibal Nayef , Thomas Breuel -- University of Kaiserslautern -- Germany
{nnayef , tmb}@iupr.com

IUPR FACHBEREICH INFORMATIK

Abstract

The ultimate goal of my research is the **reliable and efficient symbol retrieval** form large document image databases, in particular technical line drawings.

My approach to solve the retrieval problem has two main complementary directions:

First, the recognition of the symbols, both isolated recognition and in-context spotting.

Second, the off-line content analysis of line drawings, which makes us able to index the regions inside the drawings for later fast retrieval.

All my methods have been applied on standard datasets and have achieved significantly better results for spotting than the state-of-the-art approaches.

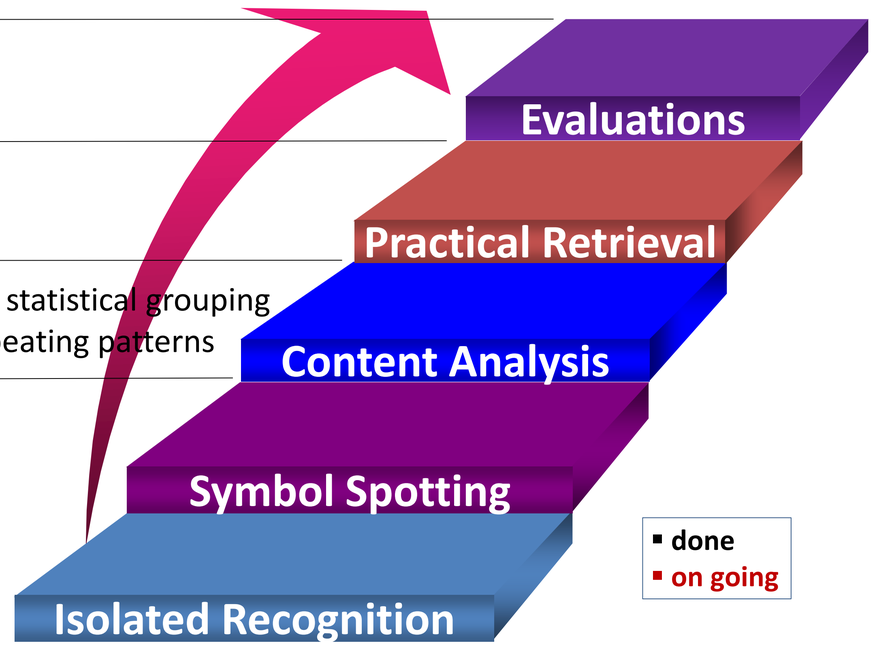
My Ph.D. Work - Overview

Effect of different parameters
Justify the work, compare to other methods

Build a symbol library
Fast matching and indexing

Identify regions of interest: Segmentation via statistical grouping
Identify patterns of interest: Clustering of repeating patterns

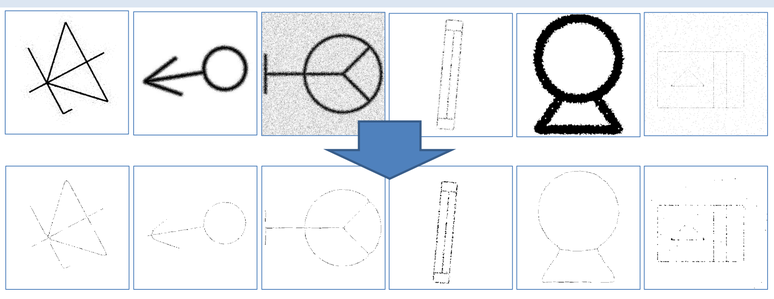
Preprocessing
Adaptive noise removal
Features extraction
Matching, Matching in-context
Post-processing
Penalizing non-matched features



I. Symbol Recognition and Spotting

A) Adaptive preprocessing [1,2]

- Automatic adaptive noise removal
- Features: pixels or line segments



B) Recognition and in-context spotting [1,3]

- Geometric matching under similarity transformations
- Branch-and-bound Search in the transformation space, align a maximal subset of features between two images

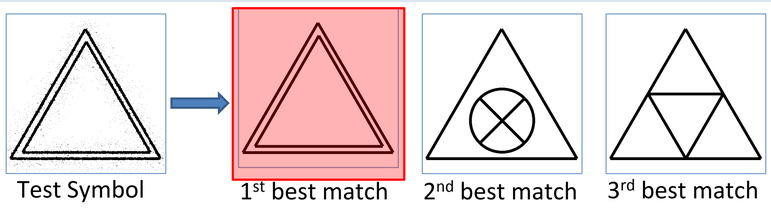


Fig: isolated symbol recognition

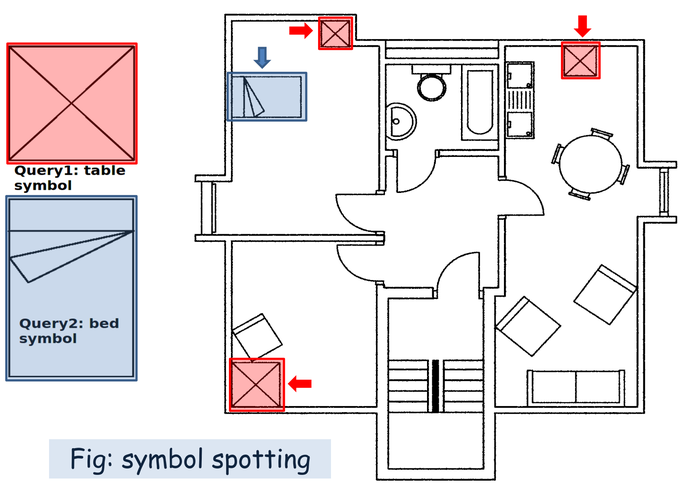


Fig: symbol spotting

Experiments & Results

Isolated Recognition* GREC'05 dataset (6000 images, all noise models, all transformations)	
Overall accuracy	90.13

Symbol Spotting*
Subset of GREC'11 training set (14 queries in 20 electrical drawings, 12 queries in 20 architectural drawings)

Data	Noise	# of instances of queried symbols	Average Recall	Average Precision
architectural	random	366	98.1	98.9
electrical	random	223	98.7	94.1

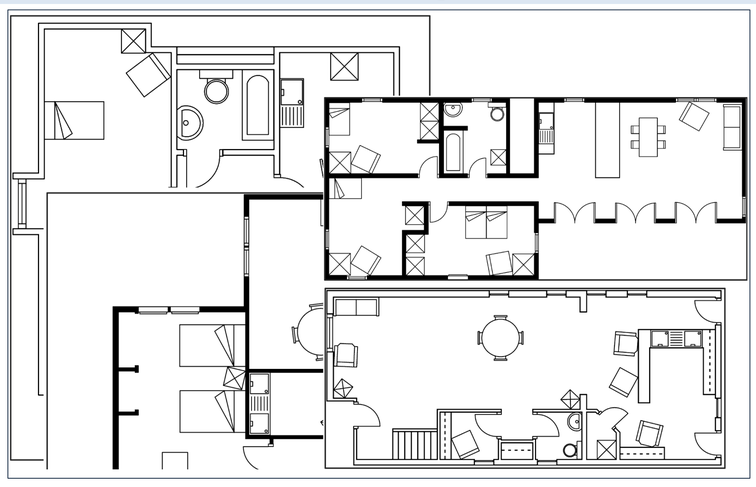
*** More results on other datasets and comparison with other methods are in [1,5]**

Participated in **GREC'11 Contest [9]** for symbol recognition and symbol spotting (all tests)

II. Symbol Retrieval by Content Analysis

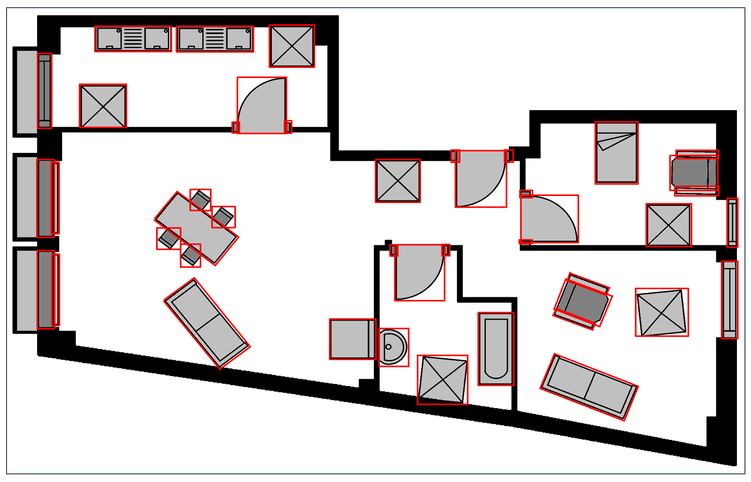
What does it do ?

From a collection of line drawings:



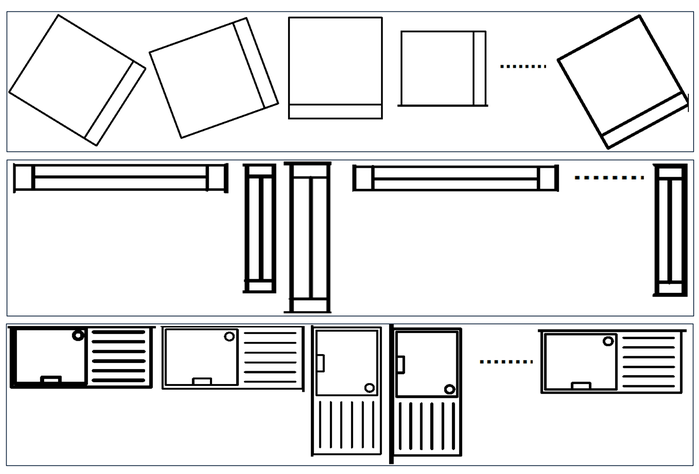
It finds – in each drawing – meaningful patterns

- using a grouping algorithm based on convex groups [5,6]
- the patterns are **symbols' parts** up to complete symbols



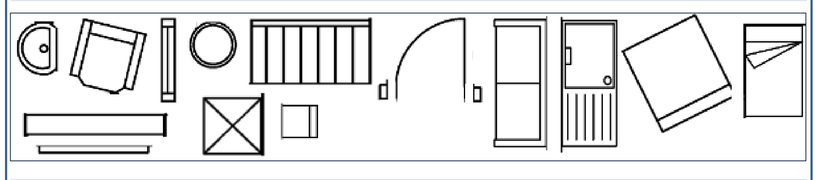
By Clustering the repeating patterns [7]

- using geometric matching [3]



It creates a symbol library from the clusters' representatives [7]:

- a very compact representation of the dataset



The symbol library will be used for fast retrieval [future]

- simply match a query symbol to the small set of symbols in the library and retrieve all matches from the matching cluster

Experiments & Results

Dataset: 300 images of complete floor plans, Public [8] , used in GREC'11 symbol spotting contest [9].

Metrics: Clustering: Recall & Precision,
Grouping: Recall for non-missing symbols,
Precision for the relevant found patterns.

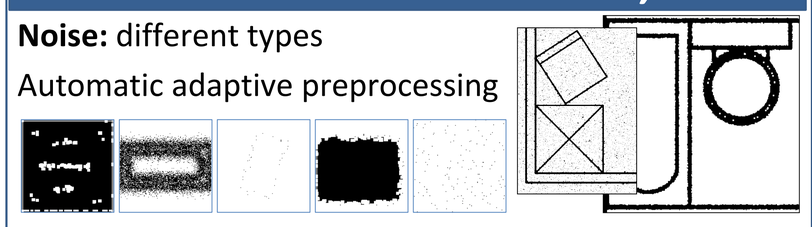
	Ground Truth	Results
No. of Patterns	12513	13780
No. of Clusters	25	30
Grouping Evaluation (offline)		
Covered Symbols	Recall (# symbols = 6987)	Precision (# patterns = 13780)
	98.8%	97.31%
Avg. time per image		22.75 sec.
Clustering Evaluation (offline)		
All Clusters*	Avg. Recall	Avg. Precision
	95%	96.5%
Avg. time per forming 1 cluster		45 min.

***Retrieval: depends on the # clusters not on the # images.**

Problems & Solutions in I, II

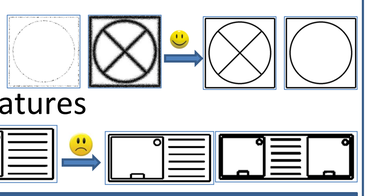
Noise: different types

Automatic adaptive preprocessing



Similar shapes

Penalizing the non-matched features inside the matching region



Take Home Messages

1) Symbol retrieval in large databases is made practical.
With very high accuracy.

2) Segmenting symbols patterns from the background has converted spotting to isolated recognition.
Using statistically justified methods .

3) Geometric matching is optimal for matching shapes under similarity transformations.
Based on well founded theories.

4) Finding repeating patterns in images leads to finding the meaningful objects parts.
Creates a compact representation of a dataset.

References

[1] N. Nayef and T. M. Breuel, "On the Use of Geometric Matching for Both: Isolated Symbol Recognition and Symbol Spotting", GREC (long paper), 2011.

[2] J. Zhang, W. Zhang, and L. Wenyin, "Adaptive noise reduction for engineering drawings based on primitives and noise assessment", in GREC, 2005, pp. 140–150.

[3] T. M. Breuel, "Implementation techniques for geometric branch-and-bound matching methods," CVIU, vol. 90, no. 3, pp. 258–294, 2003.

[4] N. Nayef and T. M. Breuel, "Graphical symbol retrieval using a branch and bound algorithm," in ICIP, 2010, pp. 2153–2156.

[5] N. Nayef and T. M. Breuel, "Statistical grouping for segmenting symbols parts from line drawings, with application to symbol spotting," accepted in ICDAR, 2011.

[6] D. W. Jacobs, "Robust and efficient detection of salient convex groups," PAMI, vol. 18, no. 1, pp. 23–37, 1996.

[7] N. Nayef and T. M. Breuel, " Building a Symbol Library From Technical Drawings By Identifying Repeating Patterns", GREC, 2011.

[8] http://mathieu.delalandre.free.fr/projects/sesyd/index.html

[9] http://iapr-tc10.univ-lr.fr/index.php/symbol-contest-2011

Contact: Nibal Nayef (nnayef@iupr.com) **https://sites.google.com/a/iupr.com/nayef**