UNIVERSITÉ DE NANTES
l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR
UMR CNRS 6597

UNSUPERVISED HANDWRITTEN GRAPHICAL SYMBOL LEARNING
-Using Minimum Description Length Principle on Relational Graph

Jinpeng LI, Harold MOUCHERE, Christian VIARD-GAUDIN

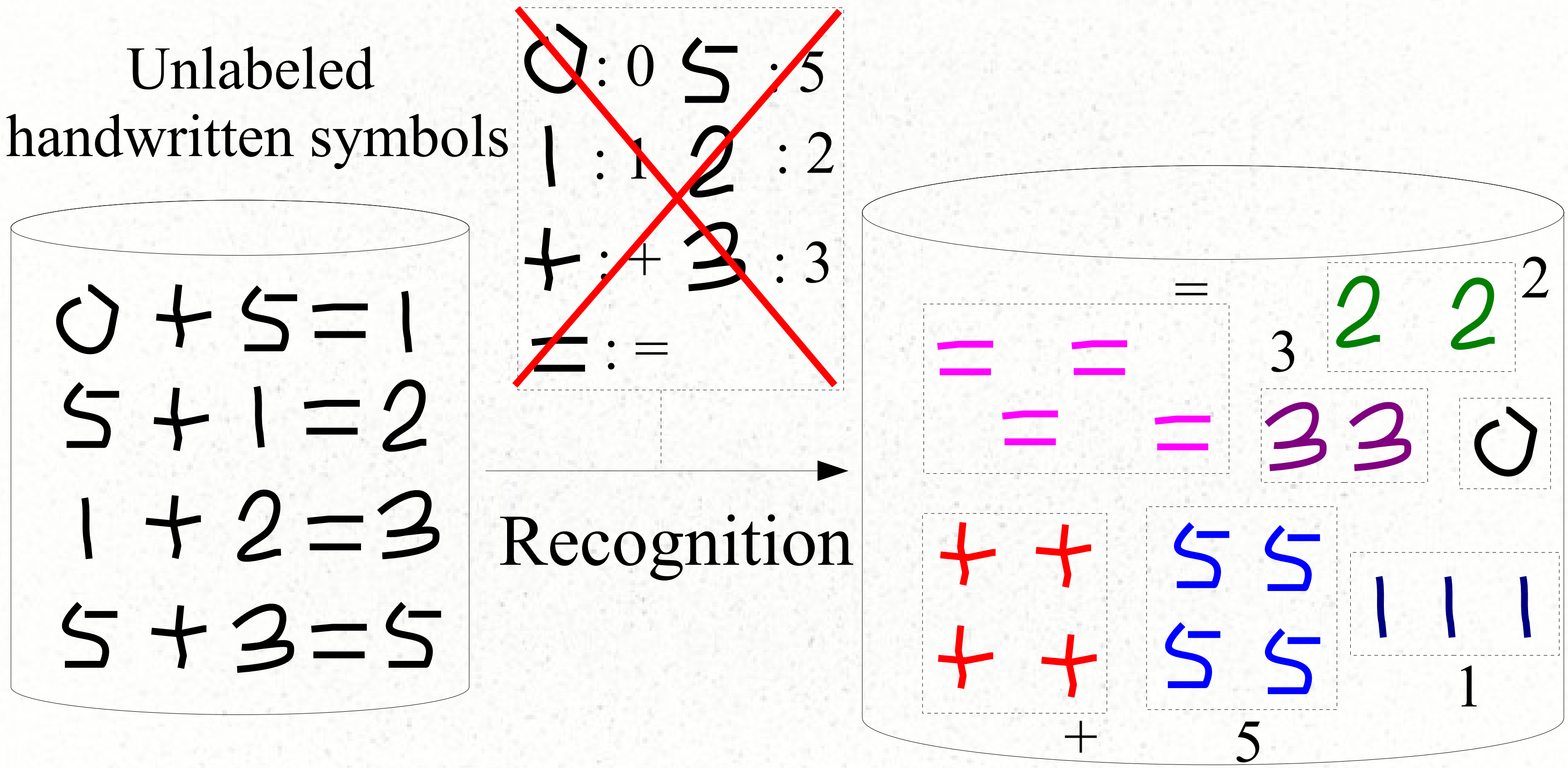
KDIR 2011

www.projet-depart.org

Outline

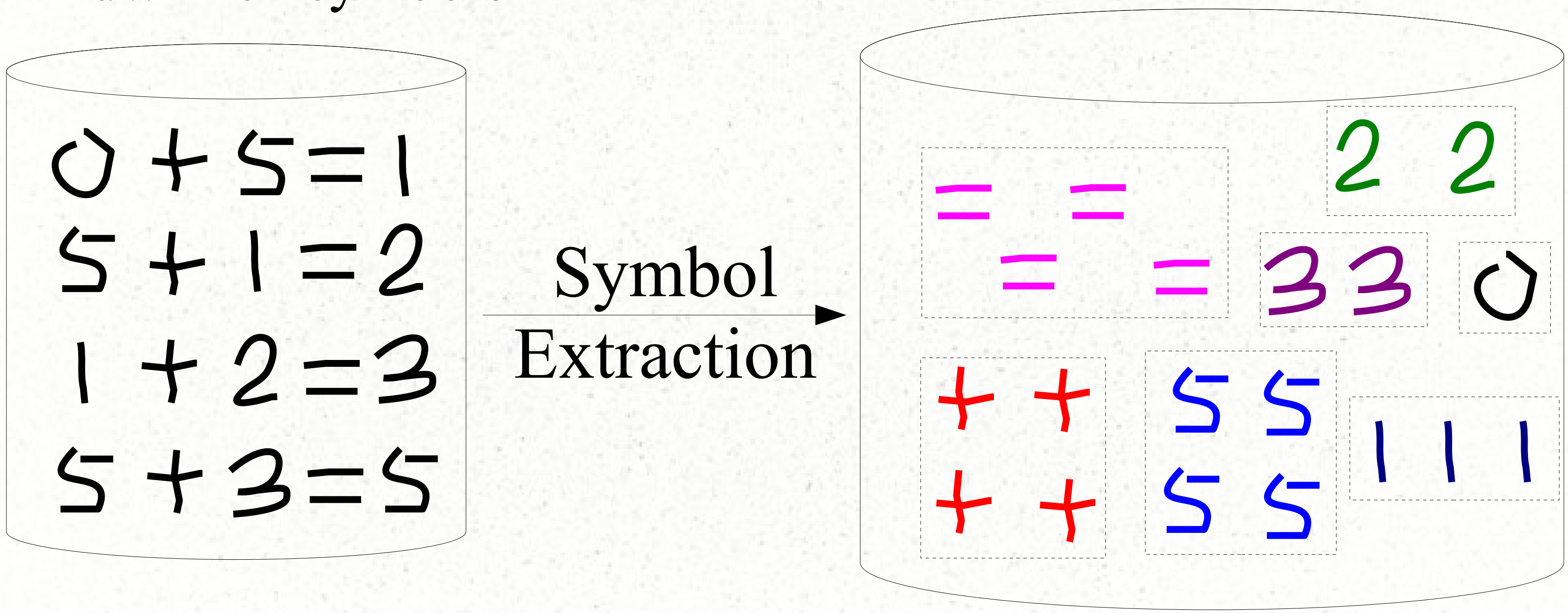
- 1. Background
- 2. Unsupervised Handwritten Graphical Symbol Learning
  - Quantization of strokes
  - Relational Graph Construction Between Strokes
  - Discover Symbols (Sub-graphs)
- 3. Experiment
- 4. Conclusion

Traditional Recognition (Background)



Annotation (Application)

Unlabeled
handwritten symbols

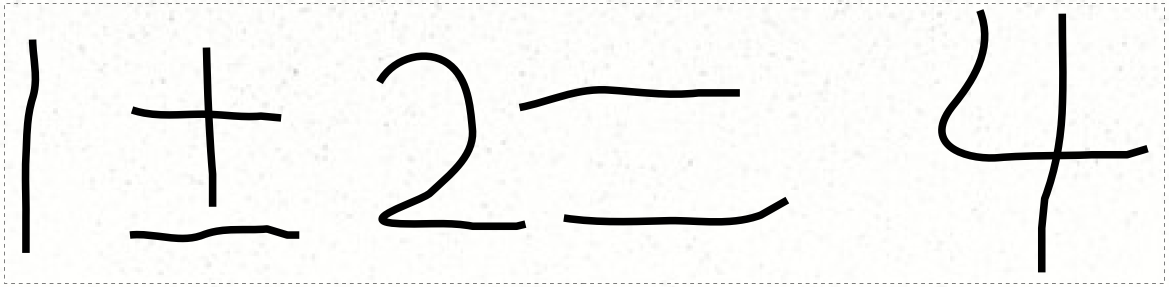


40 symbols
have to be labeled

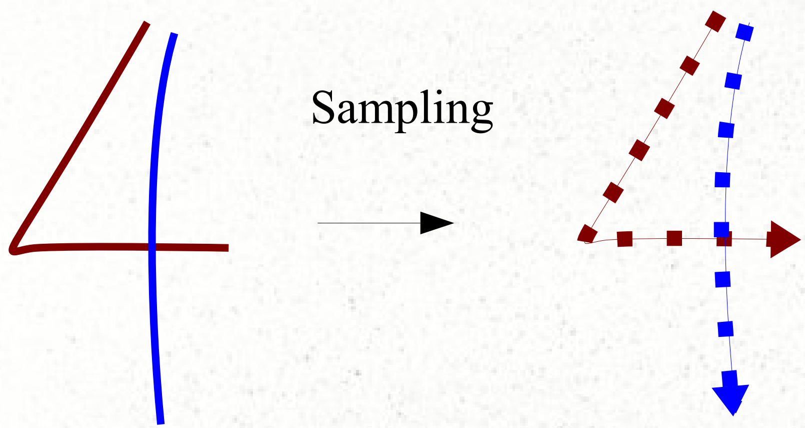
7 symbols (sets)
have to be labeled

UNSUPERVISED HANDWRITTEN GRAPHICAL SYMBOL LEARNING

As an example, we use mathematical expressions as an **unknown** graphical language.



Online handwriting

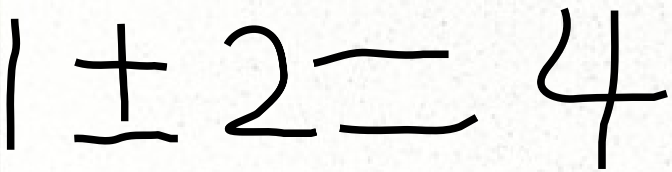


Collected data

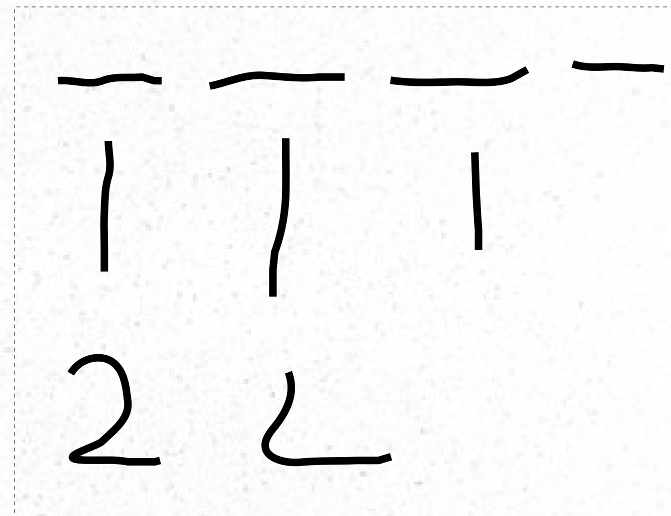
Stroke1=((x1,y1), (x2,y2),...)

Stroke2=((x1,y1), (x2,y2),...)

Unsupervised symbol learning

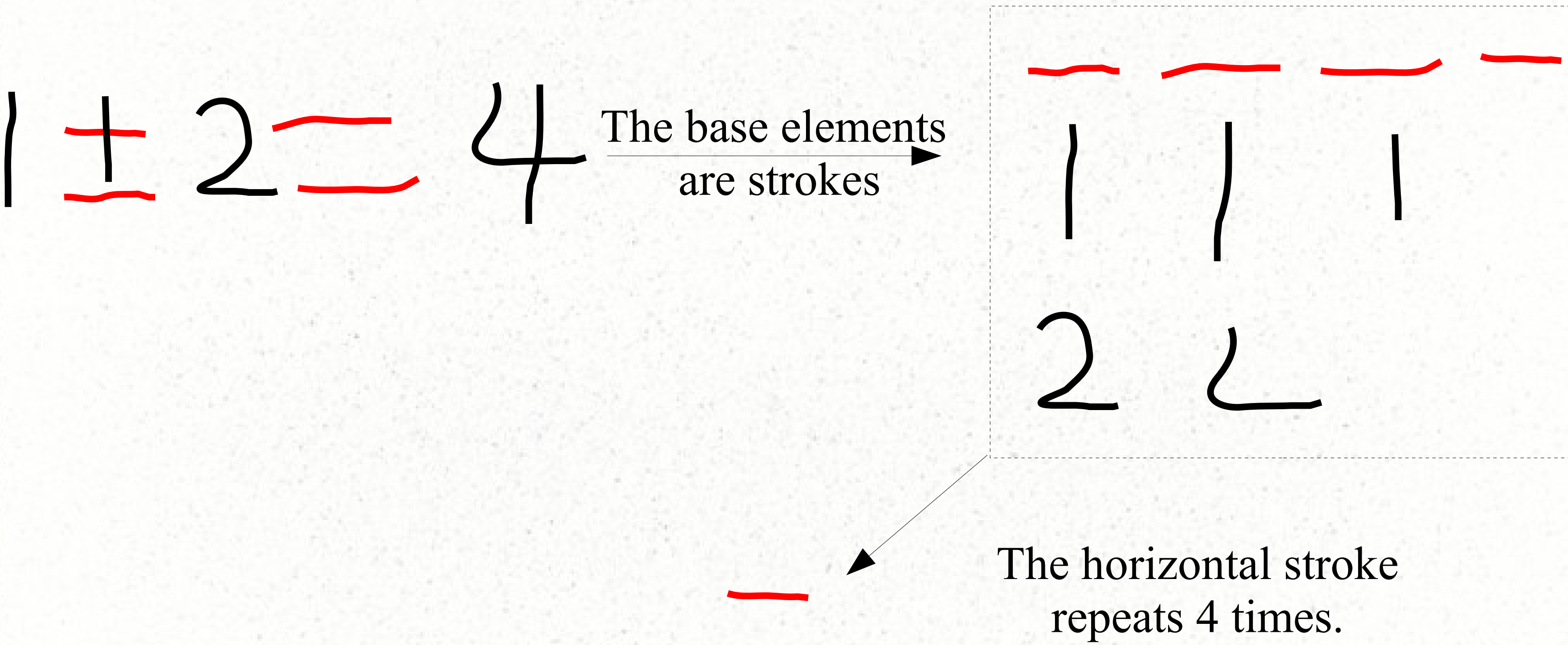


The base elements are strokes



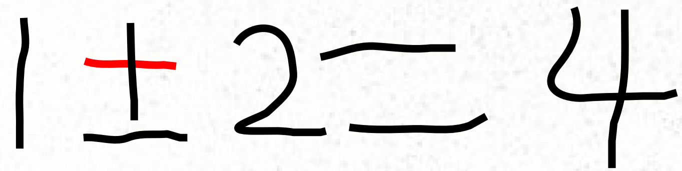
We call the frequent strokes as the **Grapheme**.

Unsupervised symbol learning

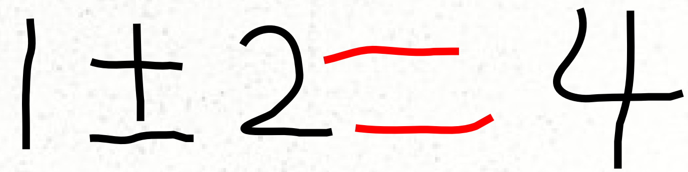


We call the frequent strokes as
the **Grapheme**.

Grapheme

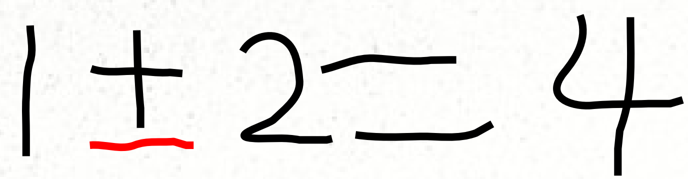


From a part of symbol "plus".



From a symbol "equal".

Where does
the horizontal stroke
come from?

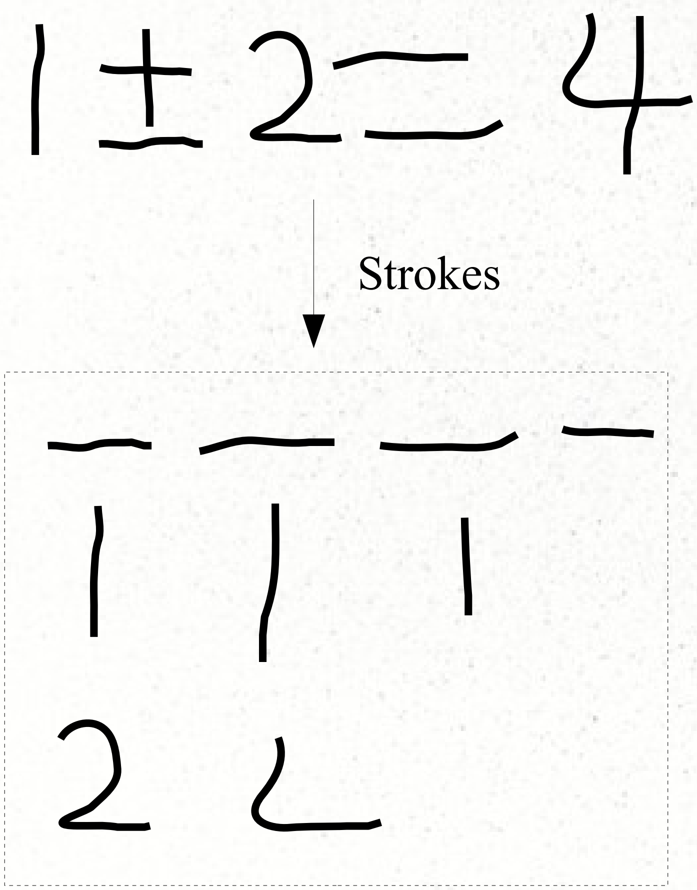


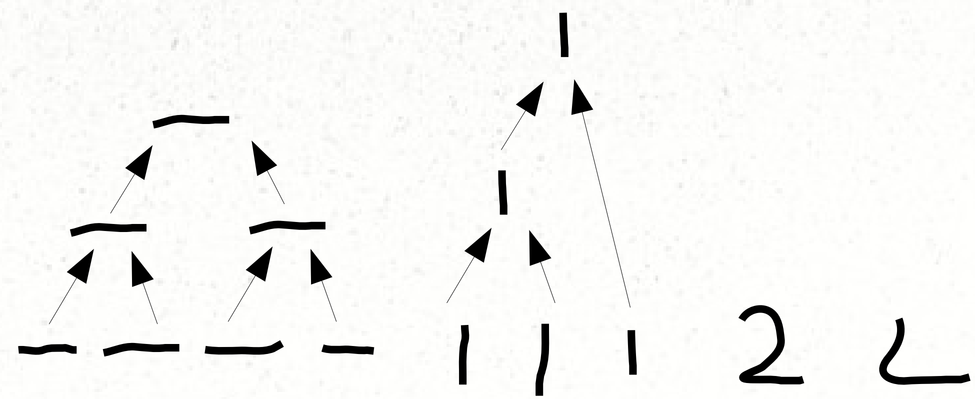
From a symbol "minus".

Outline

- 1. Background
- 2. Unsupervised Handwritten Graphical Symbol Learning
  - **Quantization of strokes**
  - Relational Graph Construction Between Strokes
  - Discover Symbols (Sub-graphs)
- 3. Experiment
- 4. Conclusion

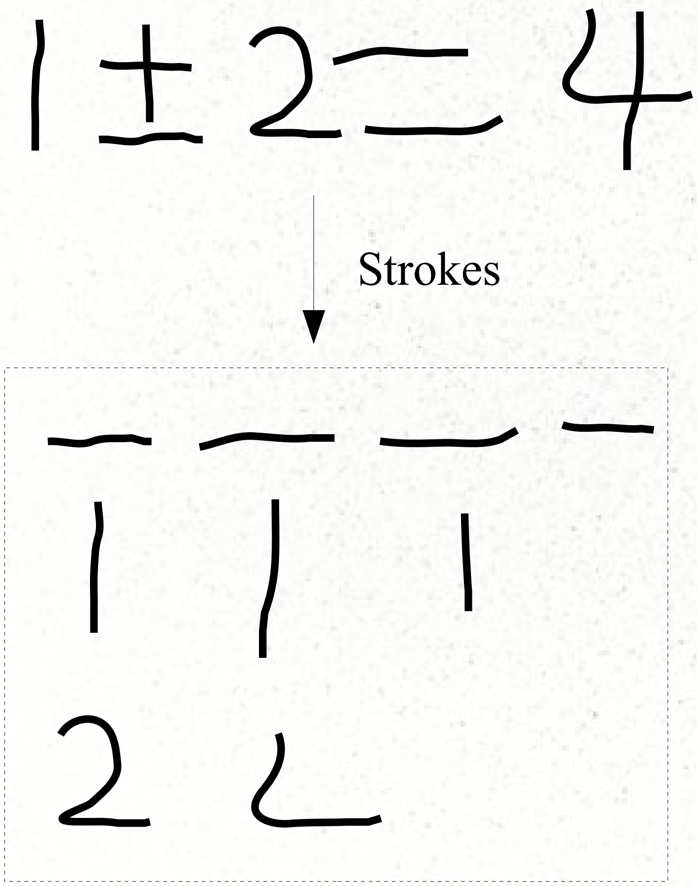
Hierarchical clustering

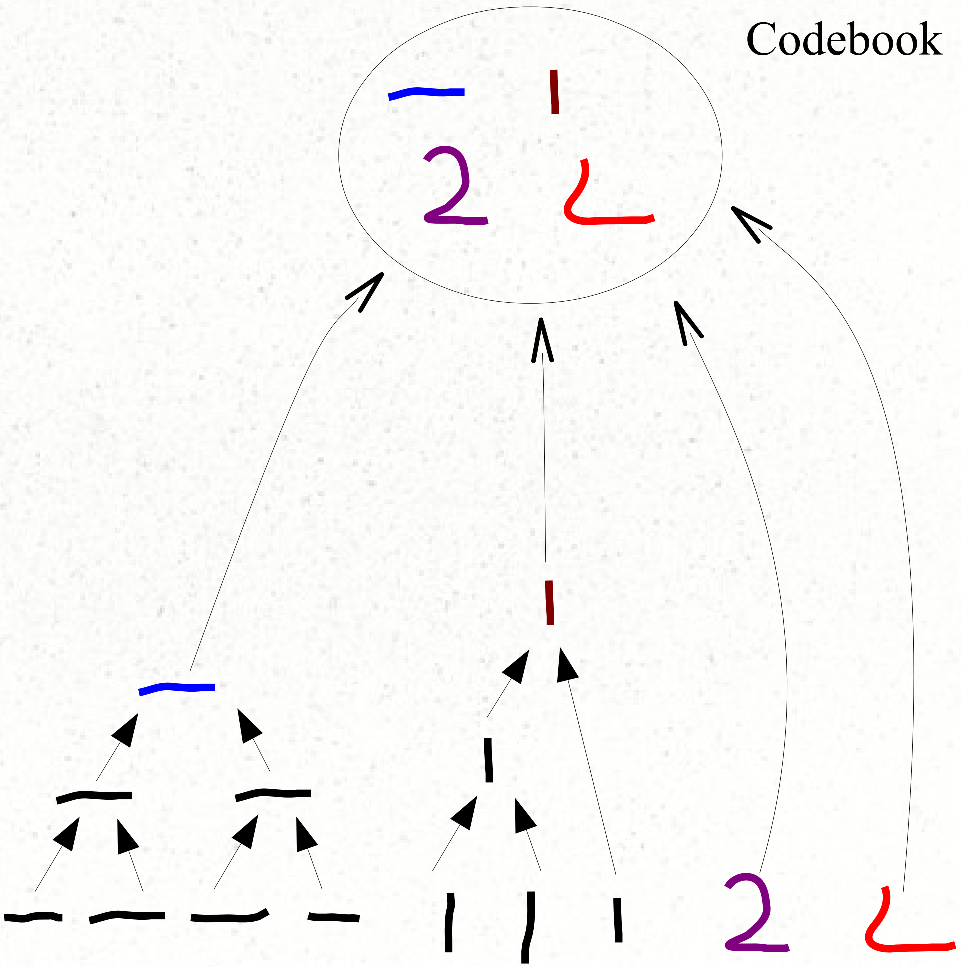




Dynamic Time Warping Distance

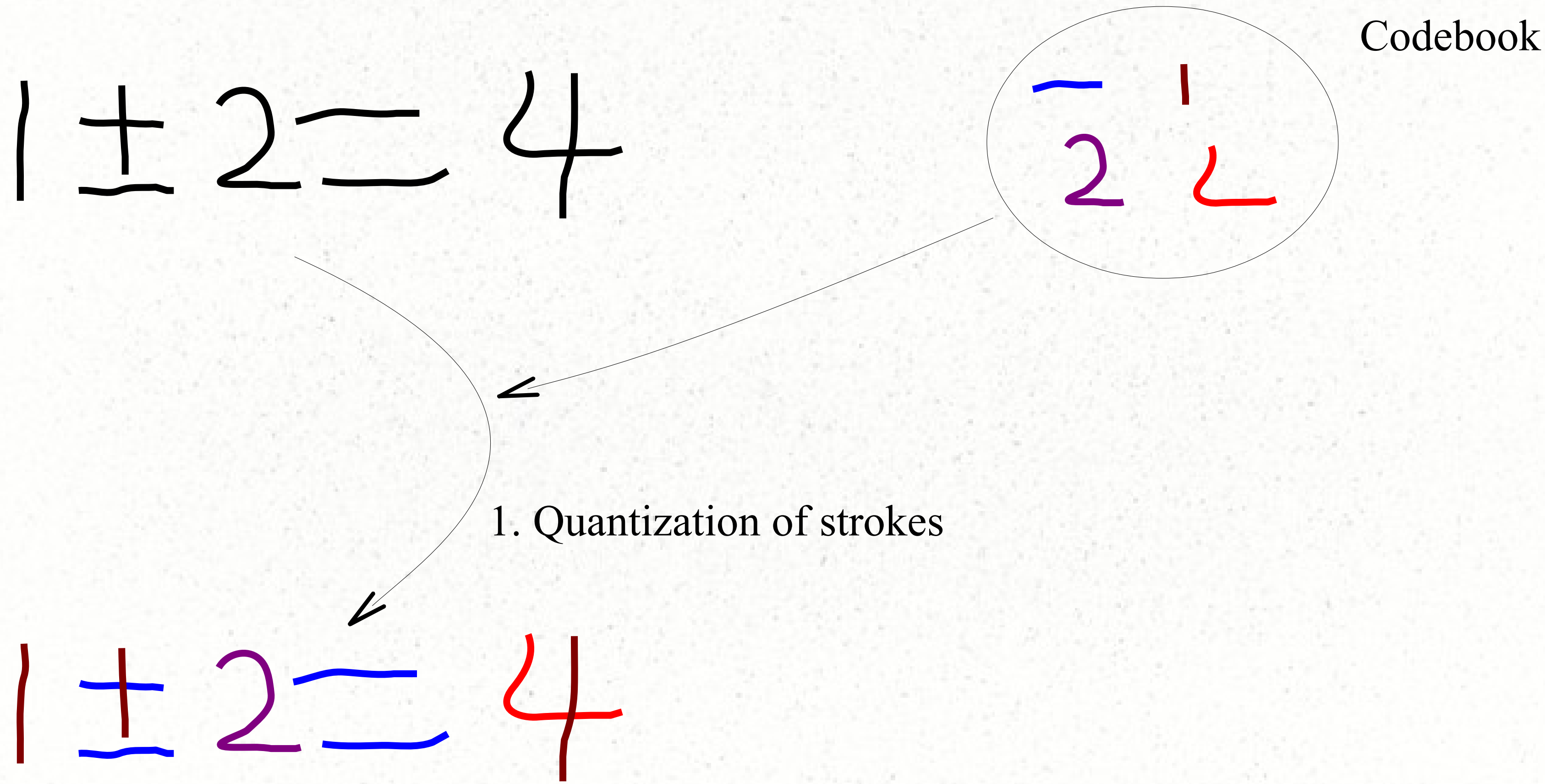
Hierarchical clustering





Dynamic Time Warping Distance

Quantization of strokes

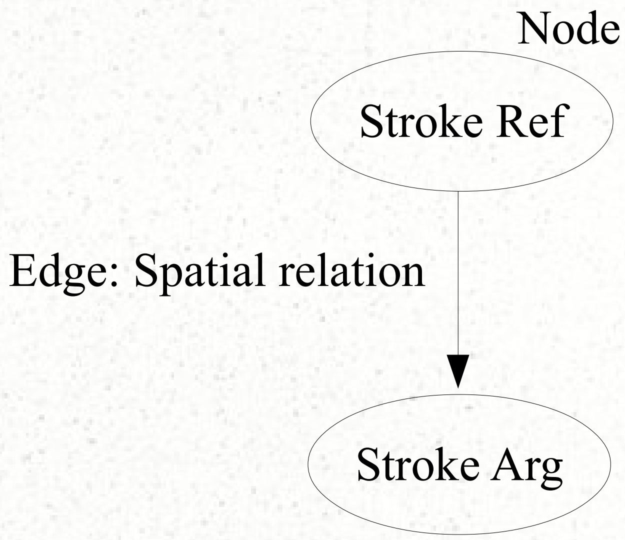


Outline

- 1. Background
- 2. Unsupervised Handwritten Graphical Symbol Learning
    - Quantization of strokes
    - **Relational Graph Construction Between Strokes**
    - Discover Symbols (Sub-graphs)
- 3. Experiment
- 4. Conclusion

Relational Graph Construction

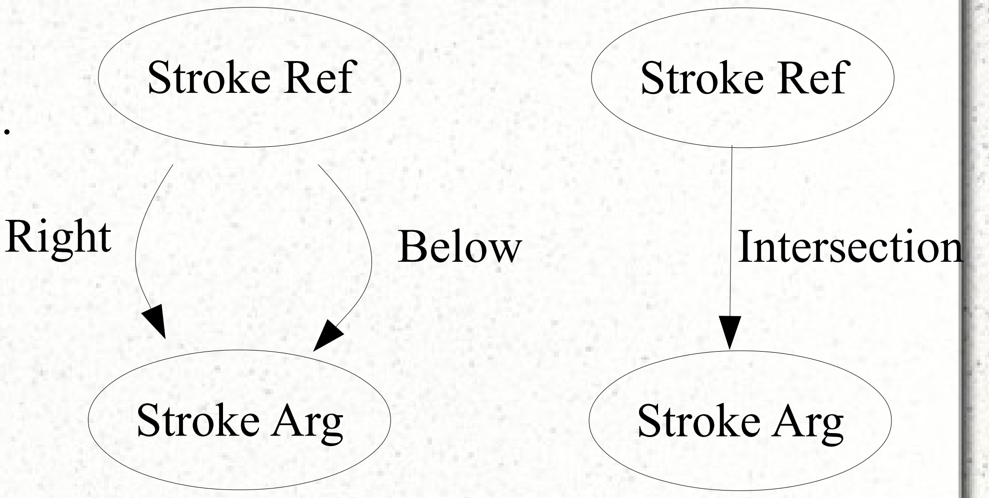
Spatial relation: from a reference stroke to an argument stroke.



Relational Graph Construction

Spatial relation: from a reference stroke to an argument stroke.

We predefine three spatial relations:
right (R), below (B), and intersection (I).



n_{str} : the number of strokes (number of nodes).

n_r : the number of different spatial relations from a reference stroke to an argument stroke.

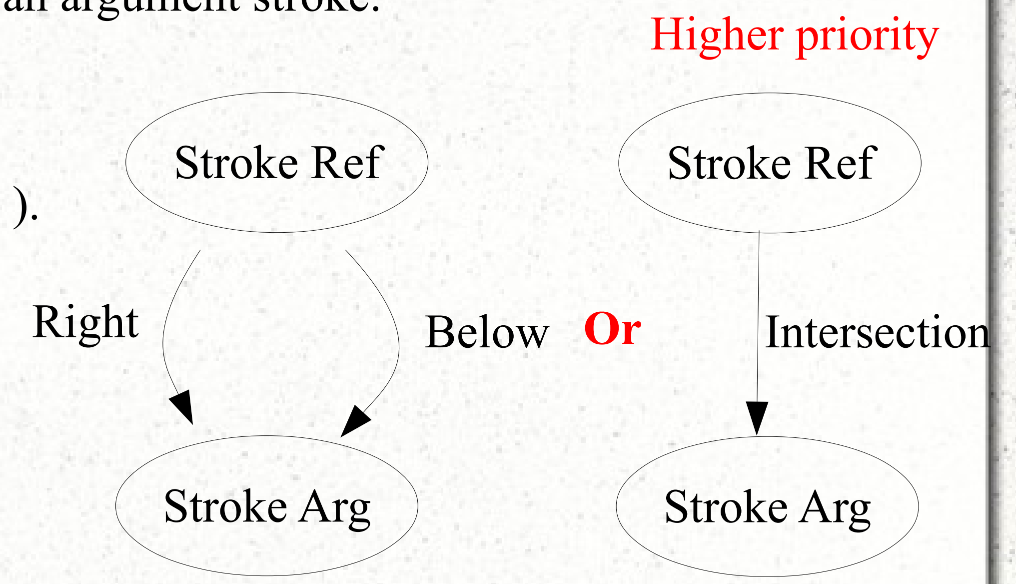
$$n_r = 3$$

Relational Graph Construction

Spatial relation: from a reference stroke to an argument stroke.

We predefine three spatial relations:
right (R), below (B), and intersection (I).

We predefine another constraint that
Directional spatial relation (R and B)
are exclusive with
Topological spatial relation (I).



n_{str} : the number of strokes (number of nodes).

n_r : the number of different spatial relations from a reference stroke to an argument stroke.

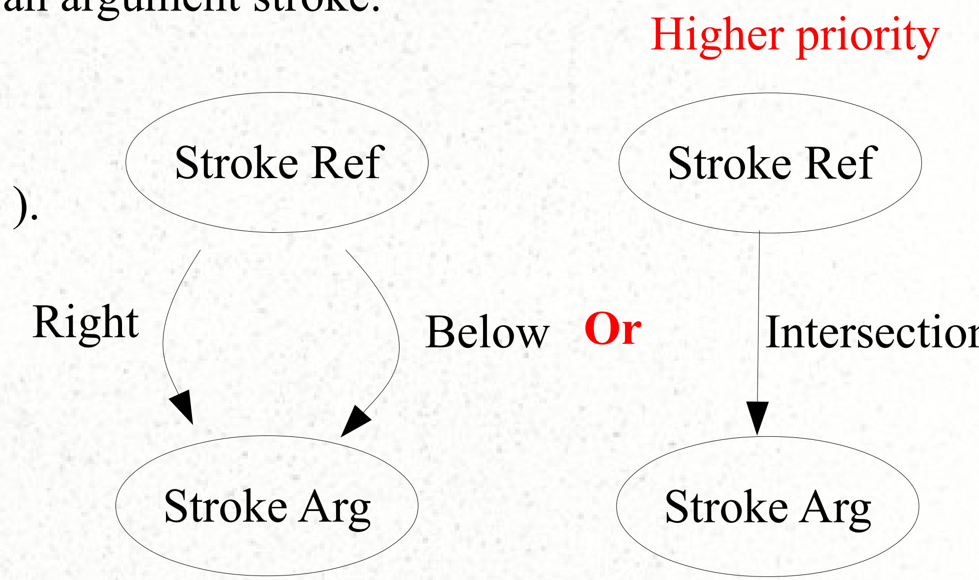
$n_r = 2$

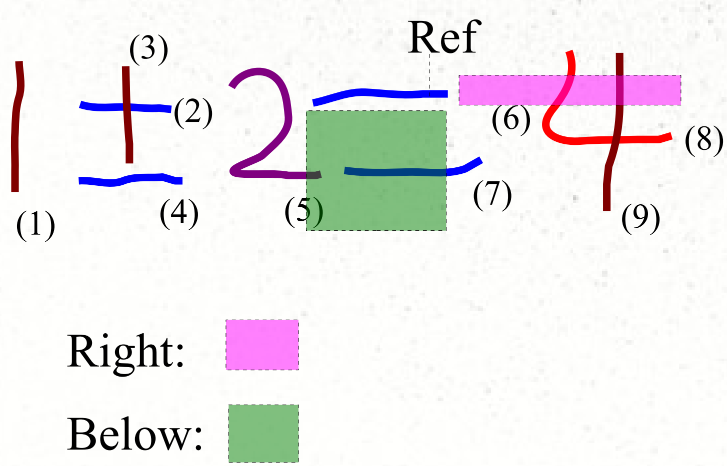
Relational Graph Construction

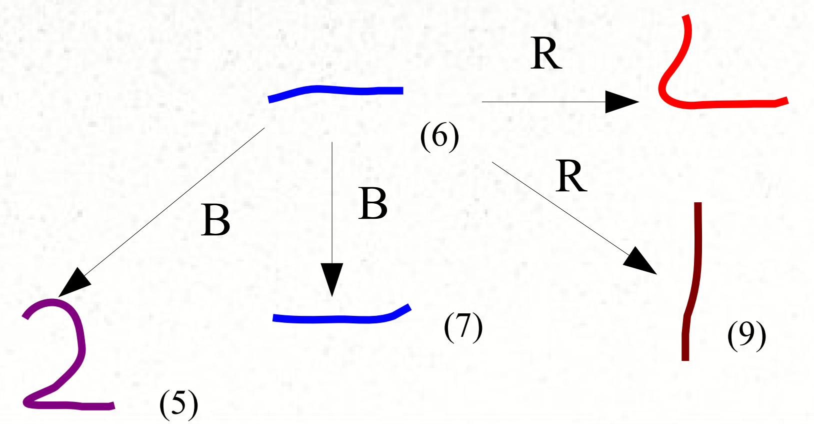
Spatial relation: from a reference stroke to an argument stroke.

We predefine three spatial relations:
right (R), below (B), and intersection (I).

We predefine another constraint that
Directional spatial relation (R and B)
are exclusive with
Topological spatial relation (I).





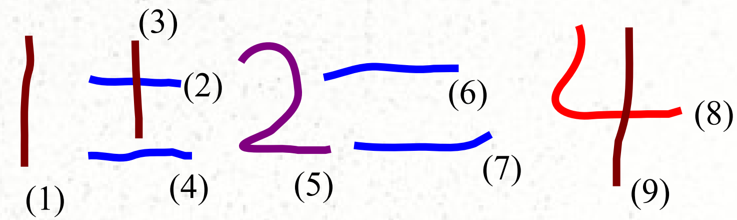


Number of edges

Spatial relation: from a reference stroke to an argument stroke.

n_r : the number of different spatial relations.

n_{str} : the number of strokes.



$n_r = 2$ $\quad n_{str} = 9$

Complete directed graph

Too many edges!

$O(n_{str}^{\ 2})$ $\quad n_r n_{str}(n_{str} - 1)$

Number of edges in graph

=144

Reduced

$n_c \le (n_{str} - 1)$

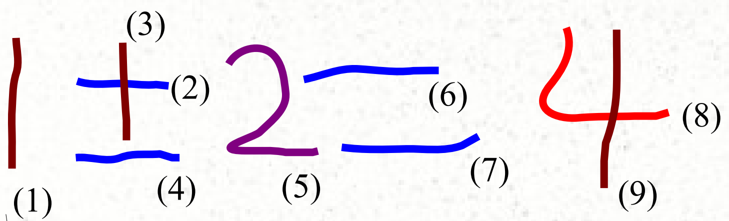
We prefer some symbols composed of the $\boxed{n_c = 2}$ closest strokes

$n_r \cdot n_{str} \cdot n_c$ $\quad 2 \cdot 9 \cdot 2 = 36$

Since we, human, have a limited perceived visual angle.

Relational Graph Construction

Spatial relation: from a reference stroke to an argument stroke.

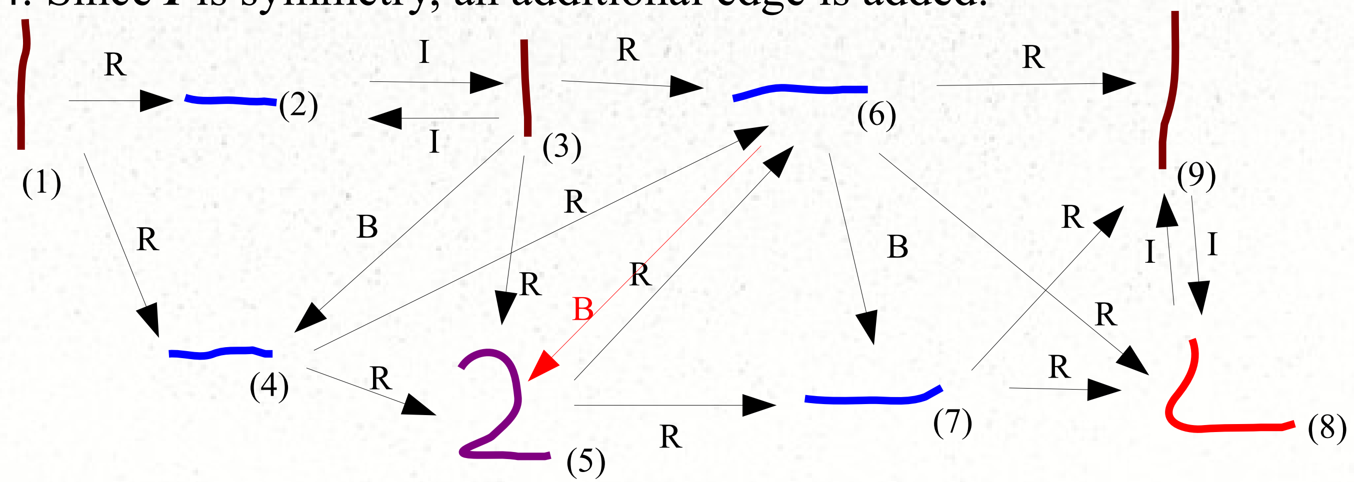


1.Start with top-left stroke.

2. Choose 2 closest strokes for each spatial relation.

3. Limit relational graph into Directed Acyclic Graph (DAG).

4. Since ***I*** is symmetry, an additional edge is added.

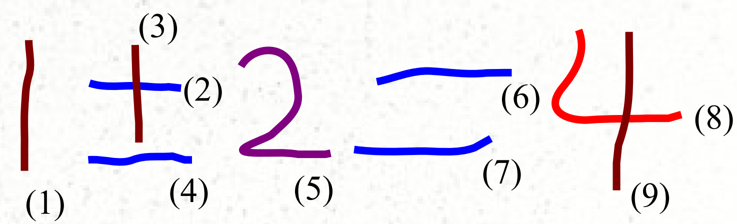


Number of edges

Spatial relation: from a reference stroke to an argument stroke.

n_r : the number of different spatial relations.

n_{str} : the number of strokes.



$n_r = 2$ $\quad n_{str} = 9$

Complete directed graph

Number of edges in graph

$$n_r n_{str}(n_{str} - 1)$$

=144

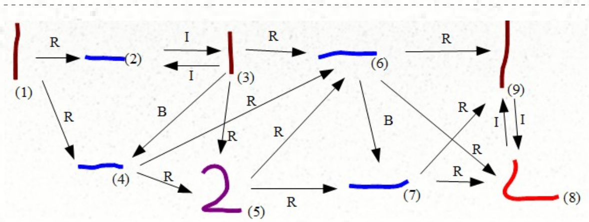
Reduced

$n_c \leq (n_{str} - 1)$

We prefer some symbols composed of the $\boxed{n_c = 2}$ closest strokes

$$n_r \cdot n_{str} \cdot n_c$$

$2 \cdot 9 \cdot 2 = 36$



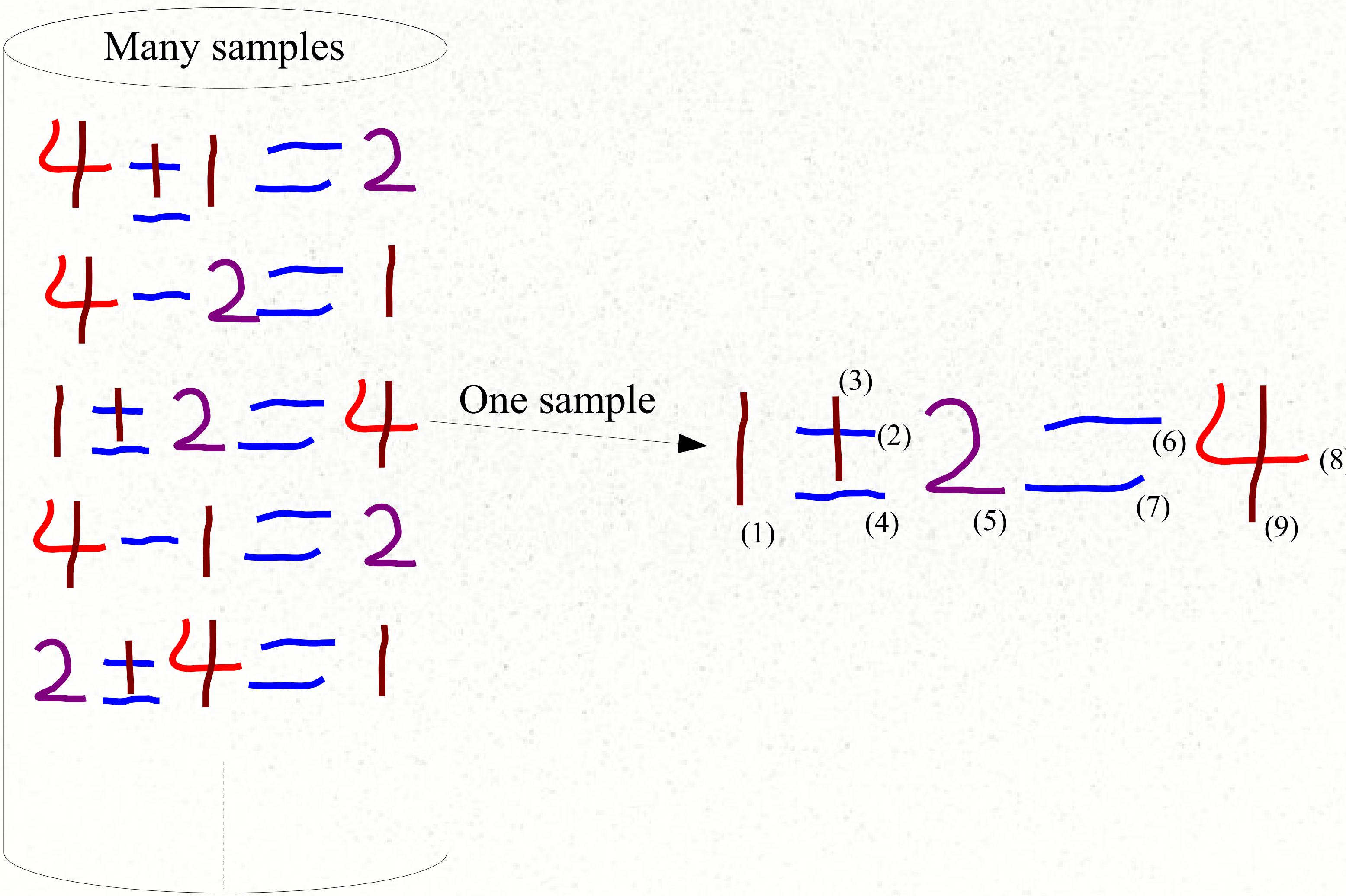
Pruning

18

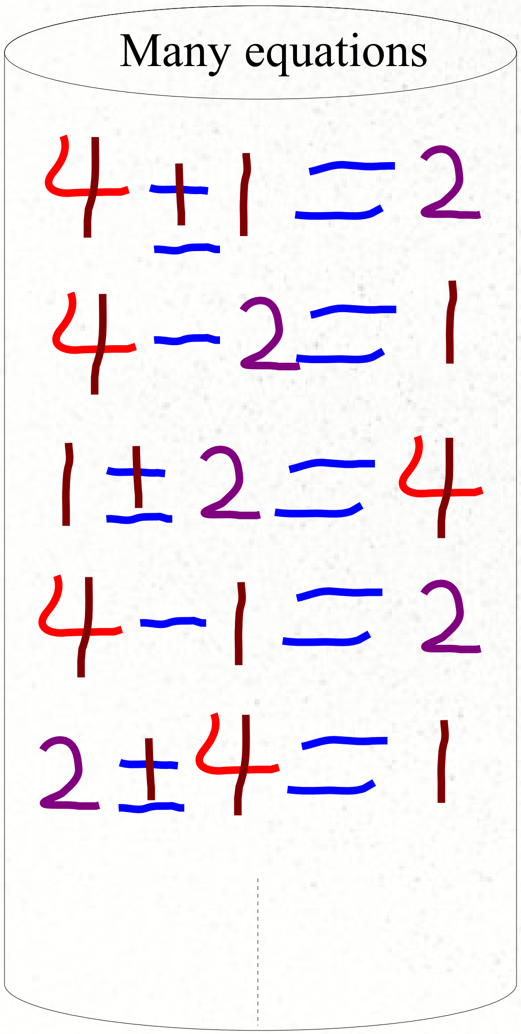
Outline

- 1. Background
- 2. Unsupervised Handwritten Graphical Symbol Learning
  - Quantization of strokes
  - Relational Graph Construction Between Strokes
  - Discover Symbols (Sub-graphs)
- 3. Experiment
- 4. Conclusion

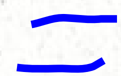
Discover Symbols (Sub-graphs)



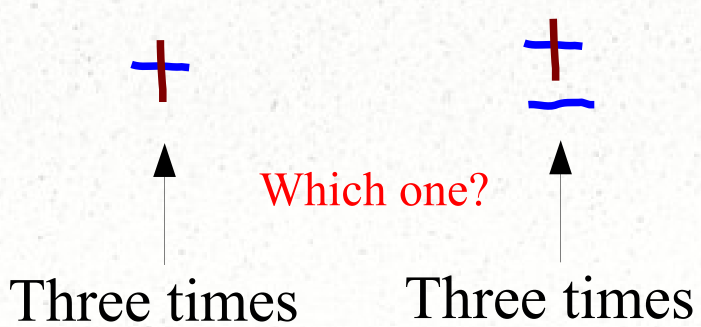
Discover Symbols (Sub-graphs)



1. We prefer the frequent patterns as the symbols.



2. Almost equally frequent pattern
   but with different numbers of strokes.



Minimum Description Length principle

Minimum Description Length (MDL) principle is involved in searching the lexical unit that leads to the best compression of data.

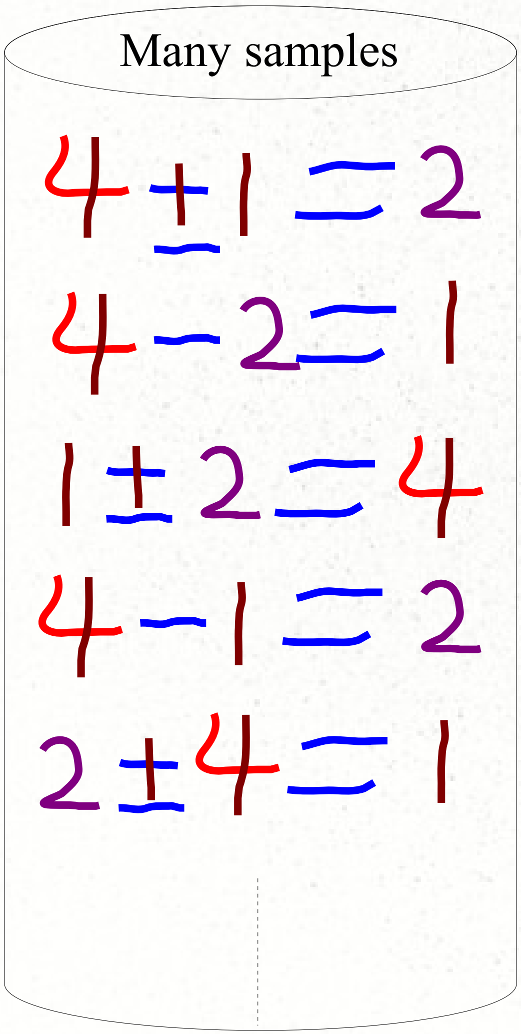
[1] describes an unsupervised language learning method using MDL principle on text corpora.

SUBDUE (SUBstructure Discovery Using Examples) uses the MDL principle to identify patterns that minimize the number of bits needed to describe the input graph after being compressed by the pattern.[2]

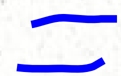
[1] Marcken, C. D., Unsupervised Language Acquisition, Massachusetts Institute of Technology, 1996

[2] Diane J. Cook and Lawrence B. Holder, http://ailab.wsu.edu/subdue/

Discover Symbols (Sub-graphs)



1. We prefer the frequent patterns as the symbols.

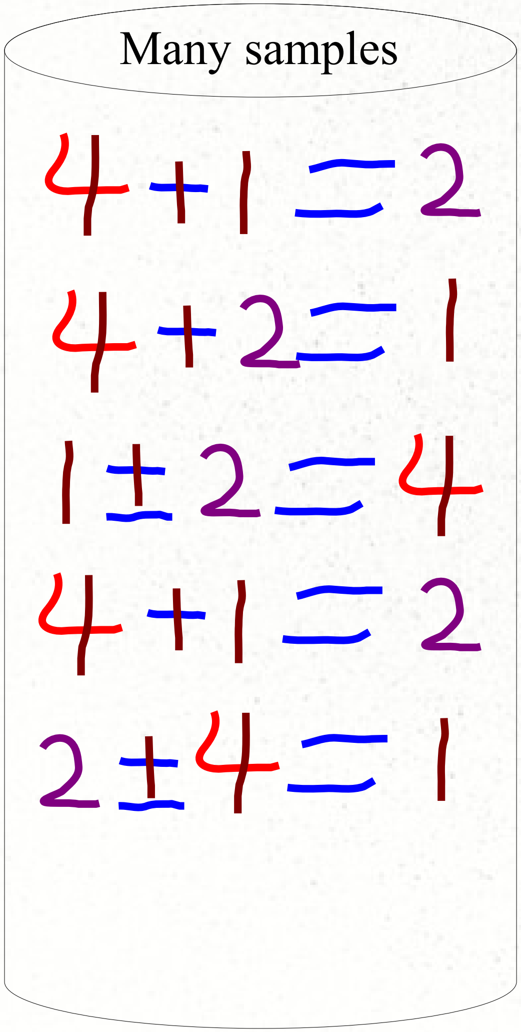


2. Almost equally frequent pattern
   but with different numbers of strokes.

Which one?

Minimum Description Length (MDL) principle

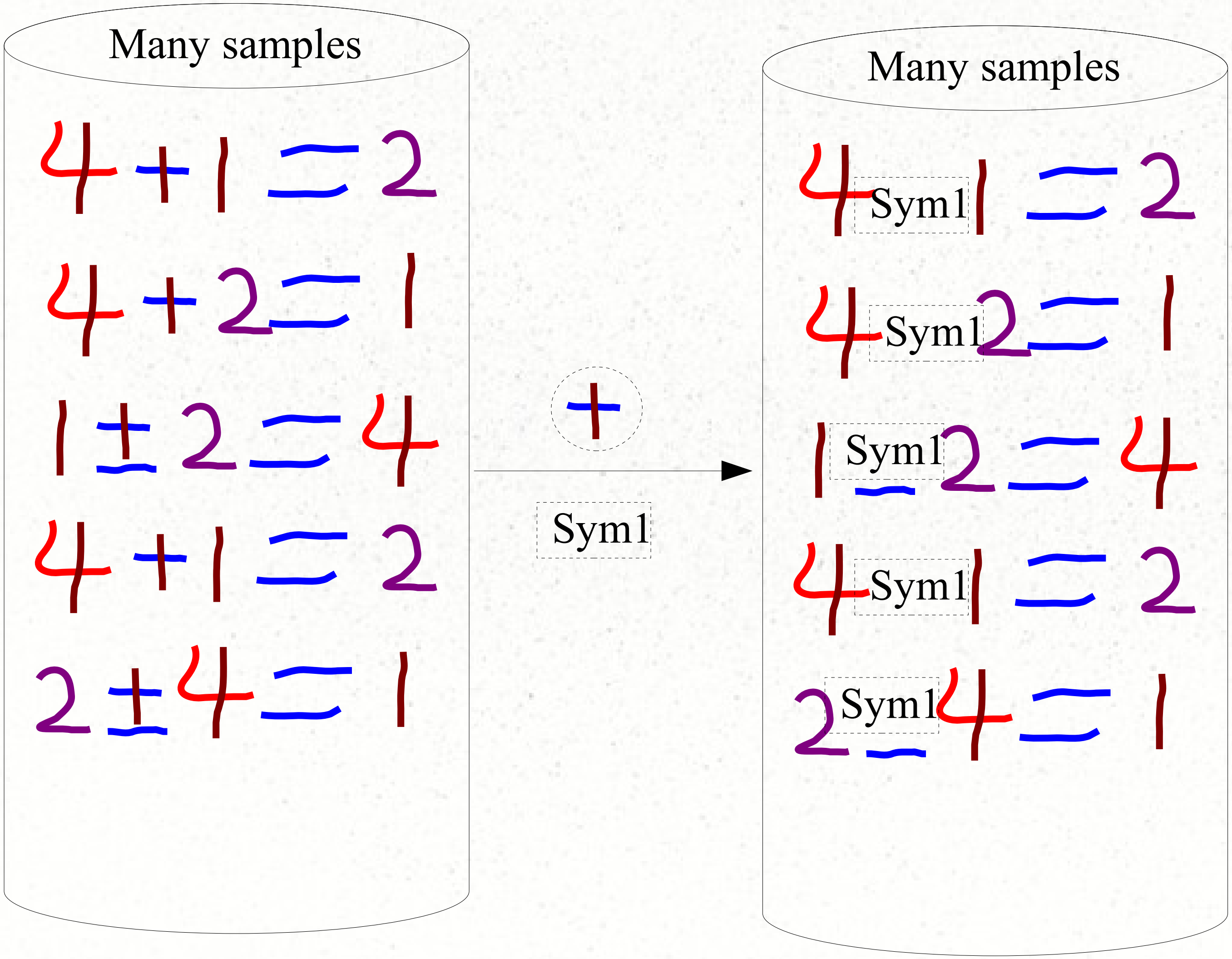
Hierarchical structure (Iterative learning)



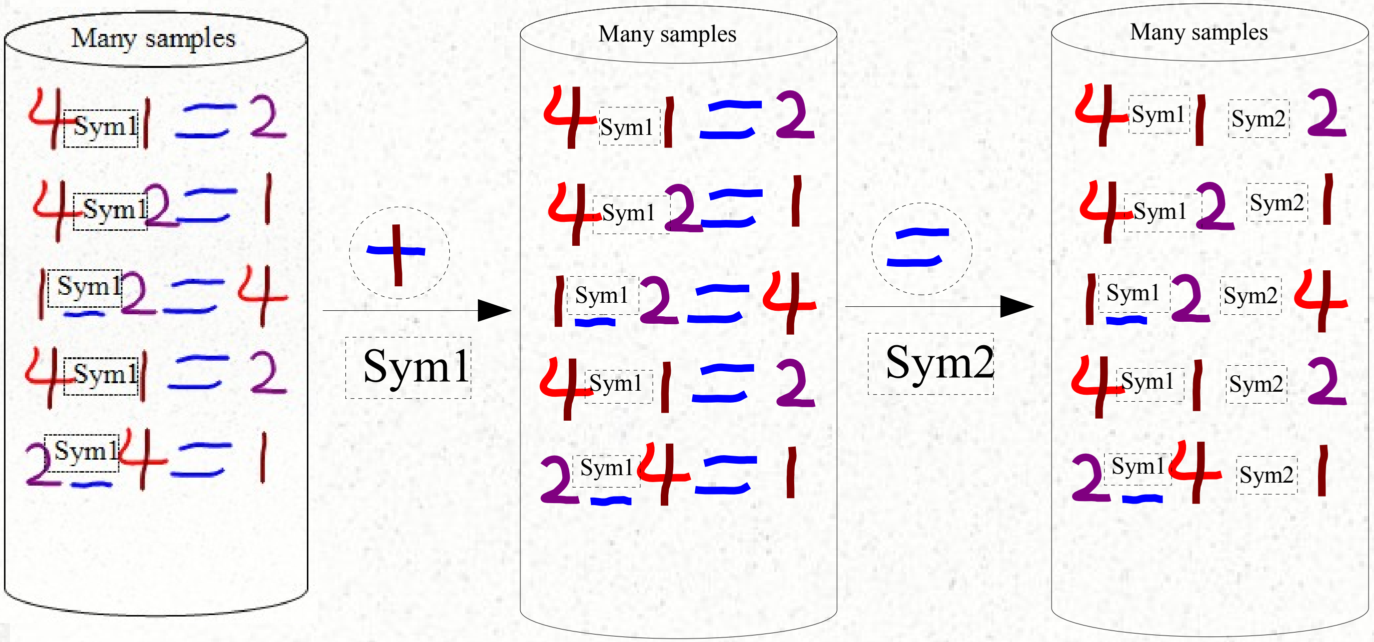
But if + is much more frequent than ± ,

we will choose the + as the symbol according to MDL principle.

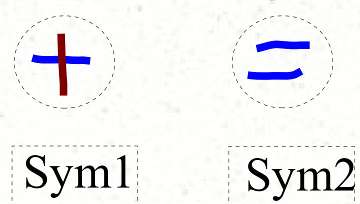
Hierarchical structure (Iterative learning)



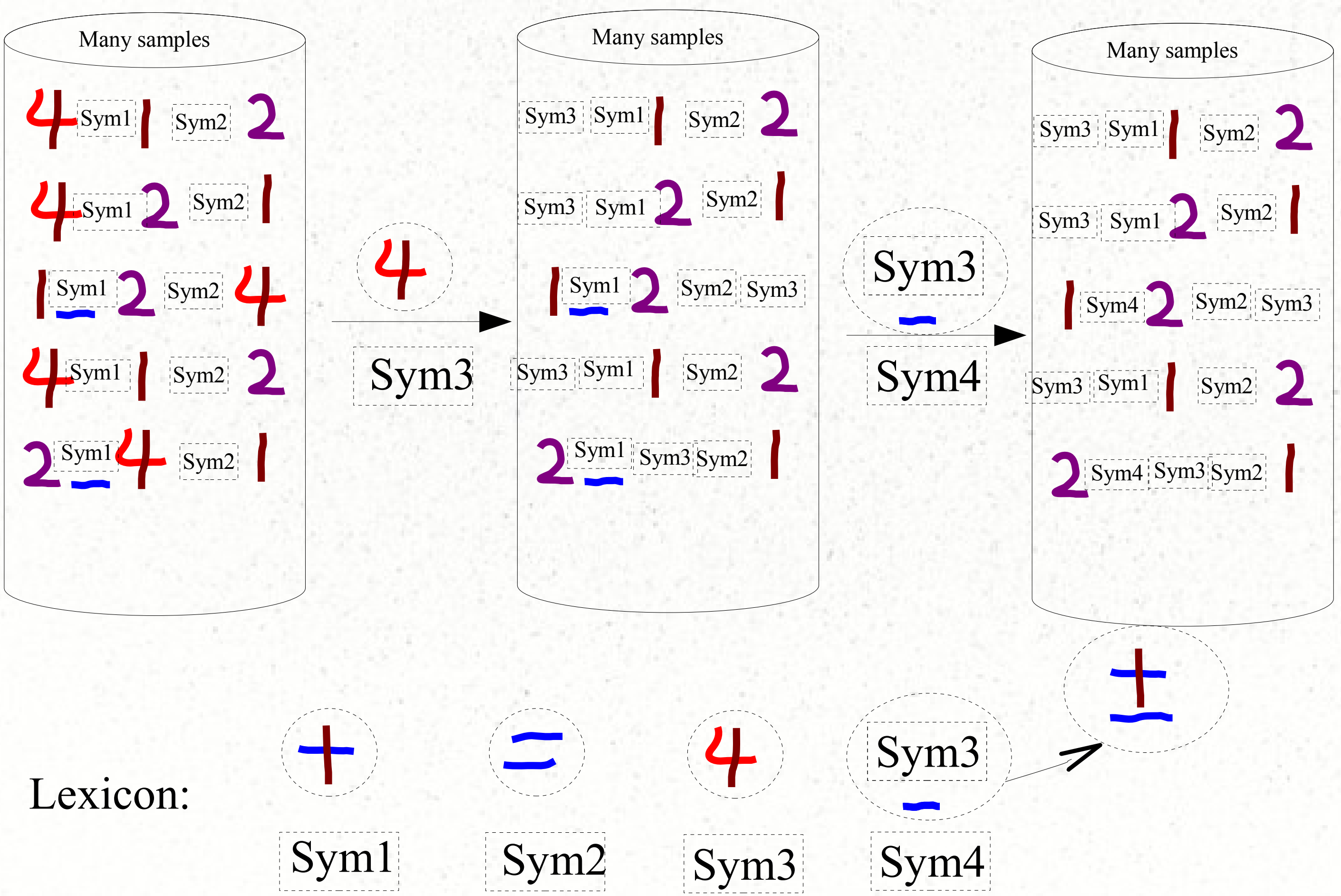
Hierarchical structure (Iterative learning)



Lexicon:



Hierarchical structure (Iterative learning)



Outline

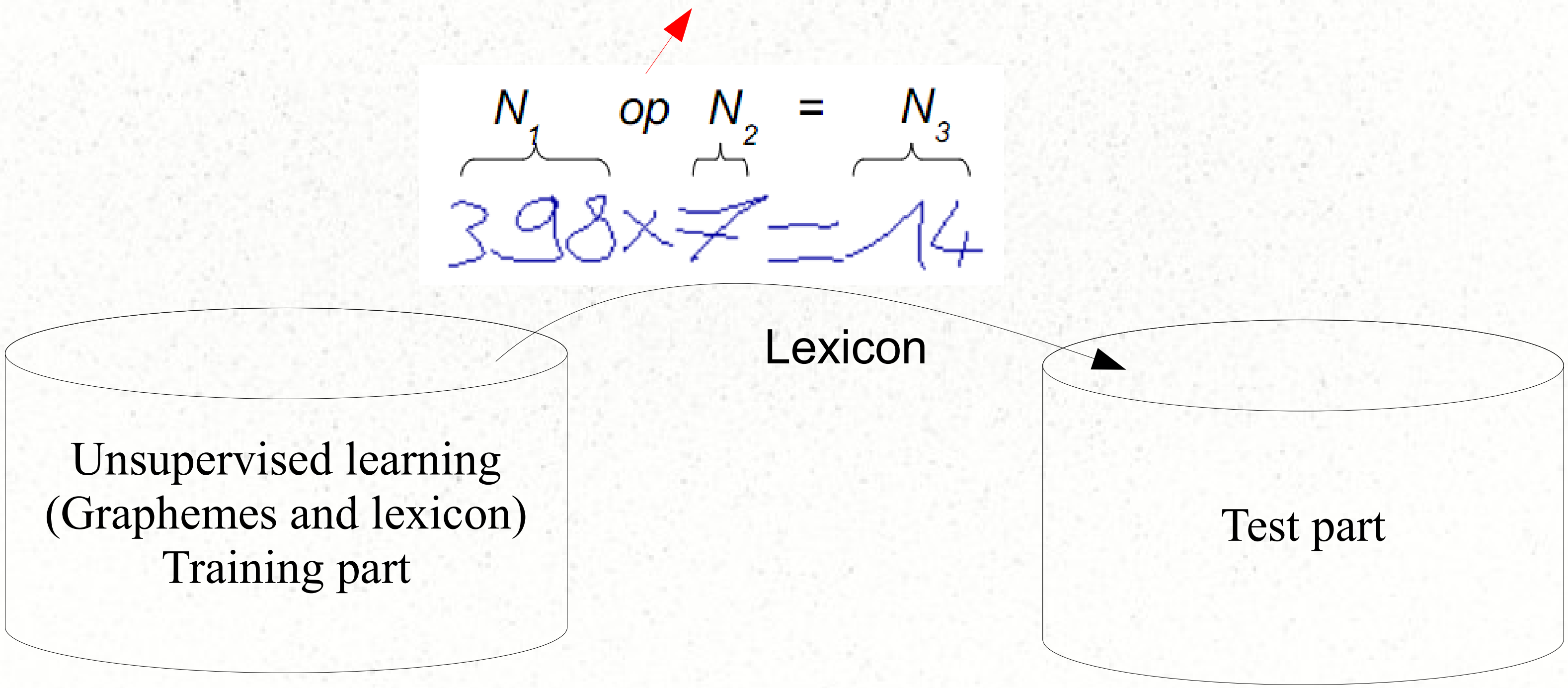
- 1. Background
- 2. Unsupervised Handwritten Graphical Symbol Learning
  - Quantization of strokes
  - Relational Graph Construction Between Strokes
  - Discover Symbols (Sub-graphs)
- 3. Experiment
- 4. Conclusion

Dataset (Experiment)

A synthetic dataset from real isolated handwritten characters.

$N_{i=\{1,2,3\}}$ is 70% of 1 digit, 20% of 2 digits and 10% of 3 digits randomly.

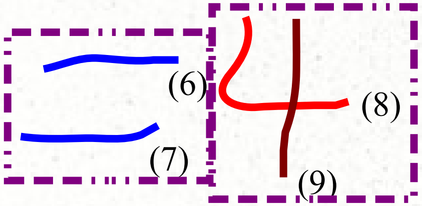
$\{0, 1, \ldots, 9\}$ $\{+, -, \times, \div\}$



5427 symbols from 180 writers

3035 symbols from 100 writers

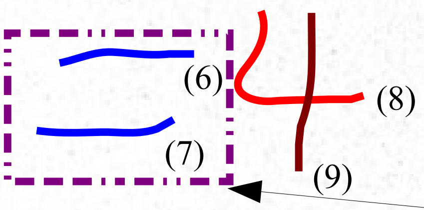
Measure (Experiment)



S(e, G):ground-truth for the expression.

$$S(e,G) = \{\boxed{\{(6),(7)\}},\boxed{\{(8),(9)\}}\}$$

Measure (Experiment)



S(e, G):ground-truth for the expression.

$$S(e,G)=\{\boxed{\{(6),(7)\}},\{(8),(9)\}\}$$

S(e, L):hierarchical segmentation using lexicon ***L***.

$$S(e,L)=\{\{(6)\},\{(7)\},\boxed{\{(6),(7)\}},\{(8)\},\{(9)\}\}$$

$$R_{\text{Recall}}=\frac{|S(e,G)\cap S(e,L)|}{|S(e,G)|}=0.5$$

We got the recall rate of ***64.3%*** for **multi-stroke** symbols (863 symbols from 1343 symbols) on the test part of our dataset.

Conclusion

- Quantization of strokes
- Construction of relational graph
- Lexicon extraction using MDL principle (SUBDUE)
- The recall rate of 64.3% (863/1343 multi-stroke symbols) is obtained.

Future work

- More complex spatial relation definition for more complex language, such as flowchart.
- Annotation assistance system for graphical symbols

Thank you for your attention.
Questions?