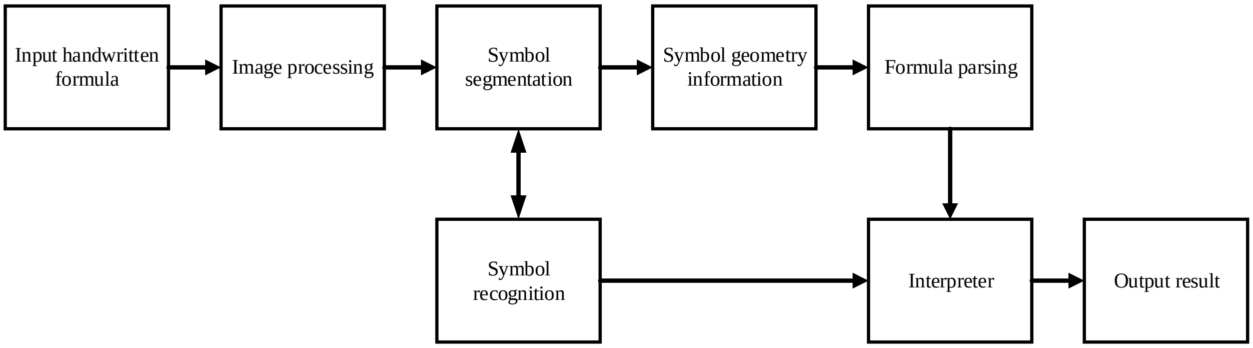
MACHINE LEARNING 2016 REPORT

PHAM DUC AN (范德安)
Student ID: 0450296
Department: Electronics Engineering
Email: anphambk@outlook.com

CHIEN-YU LIN (林建宇)
Student ID: 0450225
Department: Electronics Engineering
Email: myislin@gmail.com

Final Project: Handwritten Formula Recognition

1. Proposed System



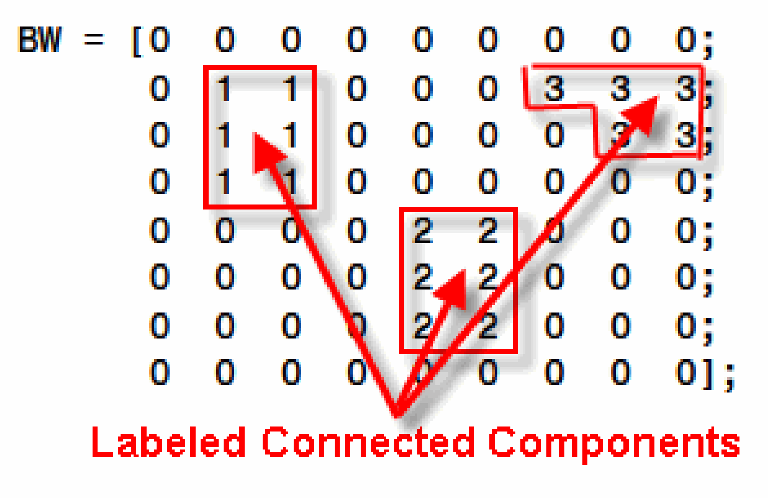
1.1. Image Processing

Handwritten formulas image only contain two colors Black and Write, so it can be convert to binary image to improve the overall performance.

Remove noises before segmentation (component with less than 100 pixels).

1.2. Symbol Segmentation

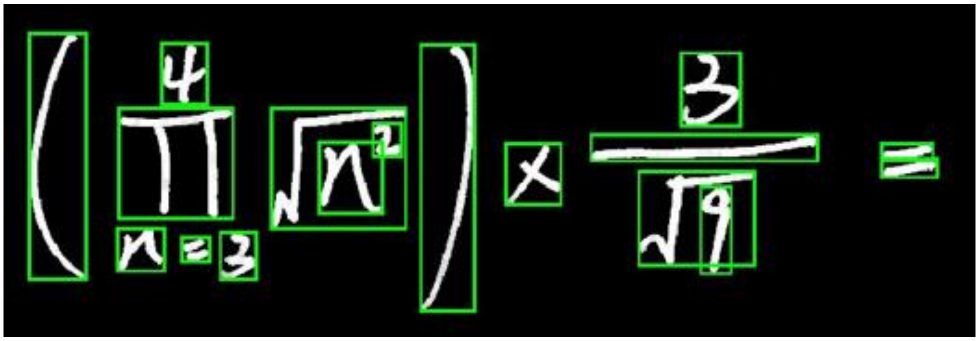
Connected Component Labeling (CCL) is the best segmentation algorithm for binary images, with high performance and high accuracy.



```
BW = [0  0  0  0  0  0  0  0  0;
      0  1  1  0  0  0  3  3  3;
      0  1  1  0  0  0  0  3  3;
      0  1  1  0  0  0  0  0  0;
      0  0  0  0  2  2  0  0  0;
      0  0  0  0  2  2  0  0  0;
      0  0  0  0  2  2  0  0  0;
      0  0  0  0  0  0  0  0  0];
```

After segmentation, we will have two information of each component: the segmented image and its geometry information. The former is use for recognition while the latter is used for formula parsing.

Matlab has already provided powerful libraries for CLL and geometry information extraction.



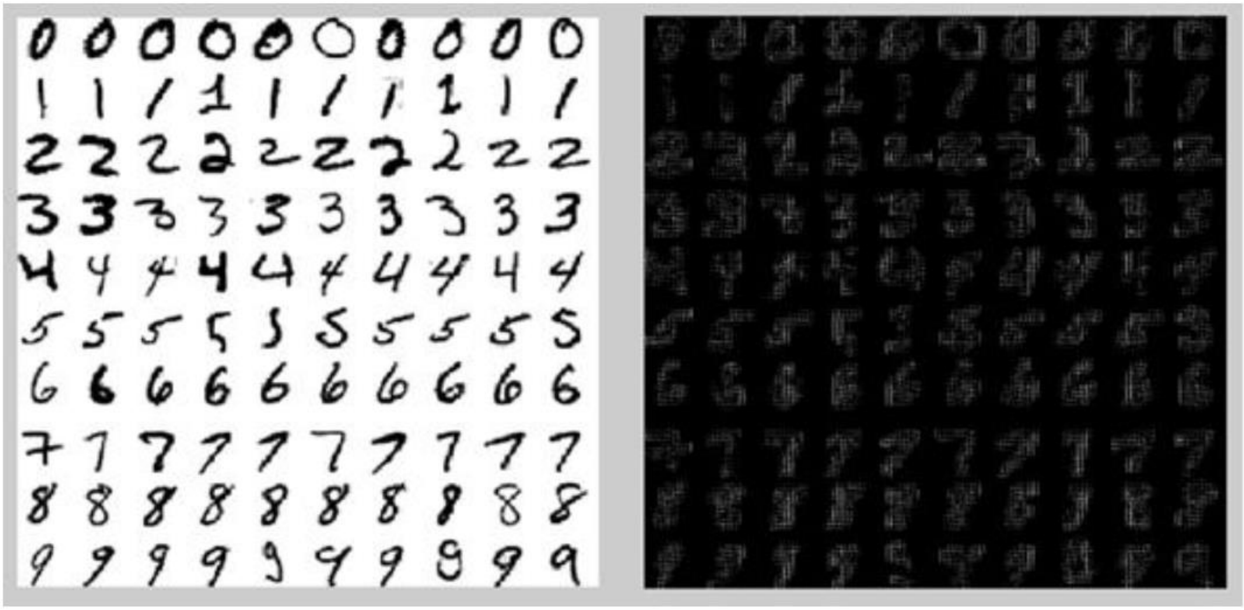
1.3. Symbol Recognition

Resize image to 40x40 to improve training process.

Use SVM as classifiers with open source library – libSVM.

Use three different classifiers with different kernels (linear, polynomial and RBF).

Beside, Histogram of Oriented Gradients (HOG) is considered when getting the features form images.



For testing phase, both classifiers with HOG and no-HOG are chosen, and the output is picked from the model which has highest probability.

Training result for v-SVM model:

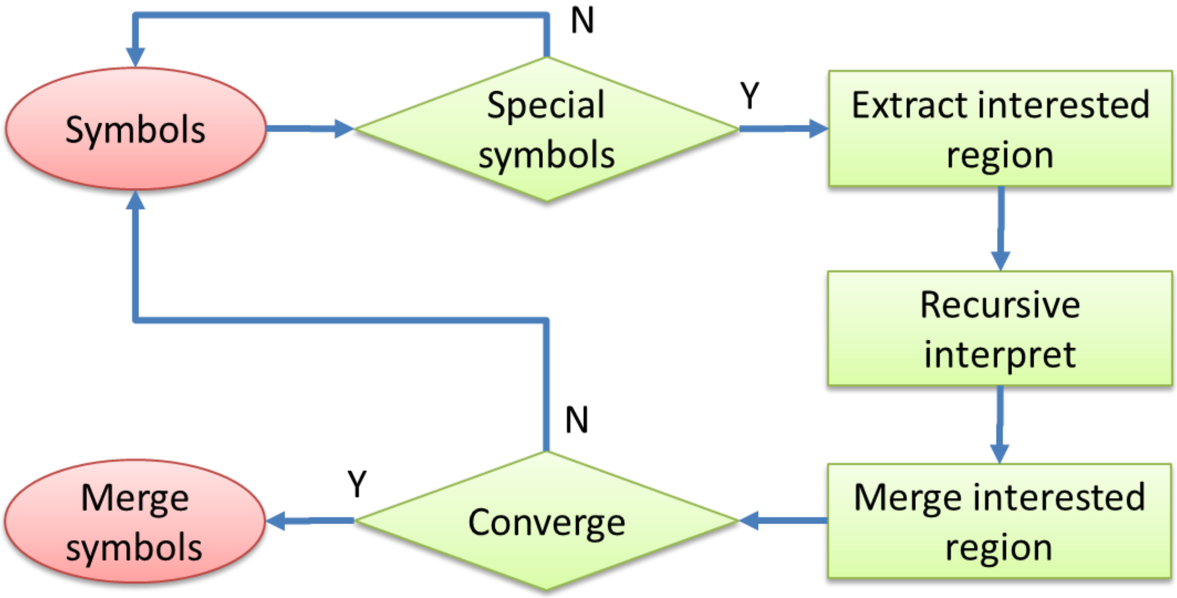
nuSVM	**Radial**	**linear**	**Poly2**	**Poly3**	**Poly4**	**Poly5**
0.001	97.597	96.6728	89.1867	15.5268	19.4085	10.9057
0.01	97.5046	96.6728	97.6895	86.3216	30.8688	16.2662
0.05	**97.6895**	**96.8577**	**97.9667**	97.2274	85.7671	56.7468
0.1	97.597	96.6728	97.3198	97.1349	88.7246	54.2514
0.25	94.9168	94.9168	96.488	96.3956	95.4713	79.9445

1.4. Formula Praising

Define interested region and find symbols that intersected with this region

- Brackets: $(\boxdot)$
- Fraction: $\frac{\boxdot}{\boxdot}$
- Power: ${\boxdot}^{\boxdot}$
- Square root: $\sqrt[\boxdot]{\boxdot}$
- Sigma: $\sum_{\boxdot}^{\boxdot}\boxdot$
- Pi: $\prod_{\boxdot}^{\boxdot}\boxdot$

The algorithm for formula parsing is as following:



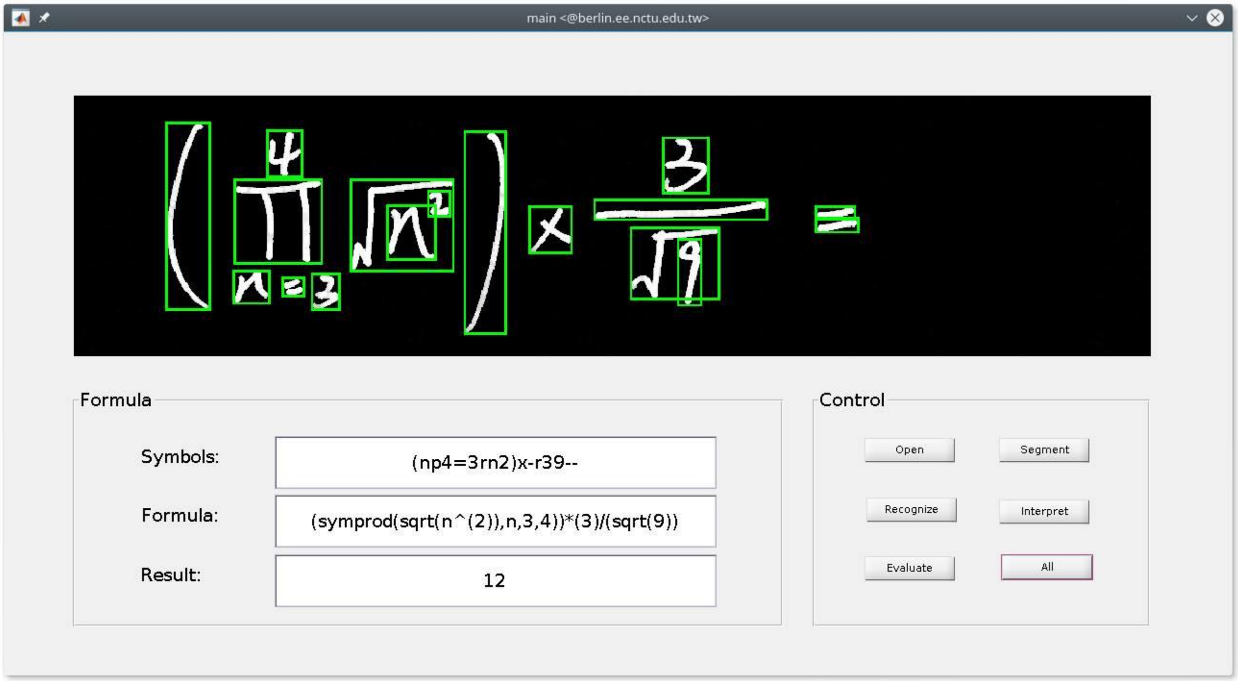
1.5. Interpreter & Evaluation

Convert to matlab code

Use matlab symbolic toolbox to deal with symbolic formulas $(m, n, \sum\quad, \prod\quad)$

2. Test Result

In order to ease the demonstration, we build a Graphic User Interface for Matlab



Symbols: (np4=3rn2)x-r39--

Formula: (symprod(sqrt(n^(2)),n,3,4))*(3)/(sqrt(9))

Result: 12

The test results:

Level	**Accuracy (%)**	**Reasons**
1	63.33	**Segment**, **Recognize**, Faulty formulas
2	76.67	**Segment**, **Recognize**, Faulty formulas
3	43.33	**Segment**, **Recognize**, Faulty formulas
4	30.00	**Segment**, **Recognize**, Faulty formulas, Interpret
5	23.33	**Segment**, **Recognize**, Faulty formulas, Wrong formulas

The main reasons for wrong output result are segmentation and recognition. There are some faulty and wrong formulas that also cause the wrong output.

3. Conclusion and Discussions

3.1. Limitations

CCL cannot segment some symbols which accidentally connected together

There is a small ratio of misclassification

Formula parsing is limited

Cannot correct wrong formulas

3.2. Proposed Solutions

Use feedback system to improve the segmentation process: if the symbol has low probability, it will be sent back to segmentation block. A new segmentation algorithm will be applied to solve this case

- Apply the symbols filter to the image
- Calculate the covariance between the filter pattern and the processing region and chose the highest one