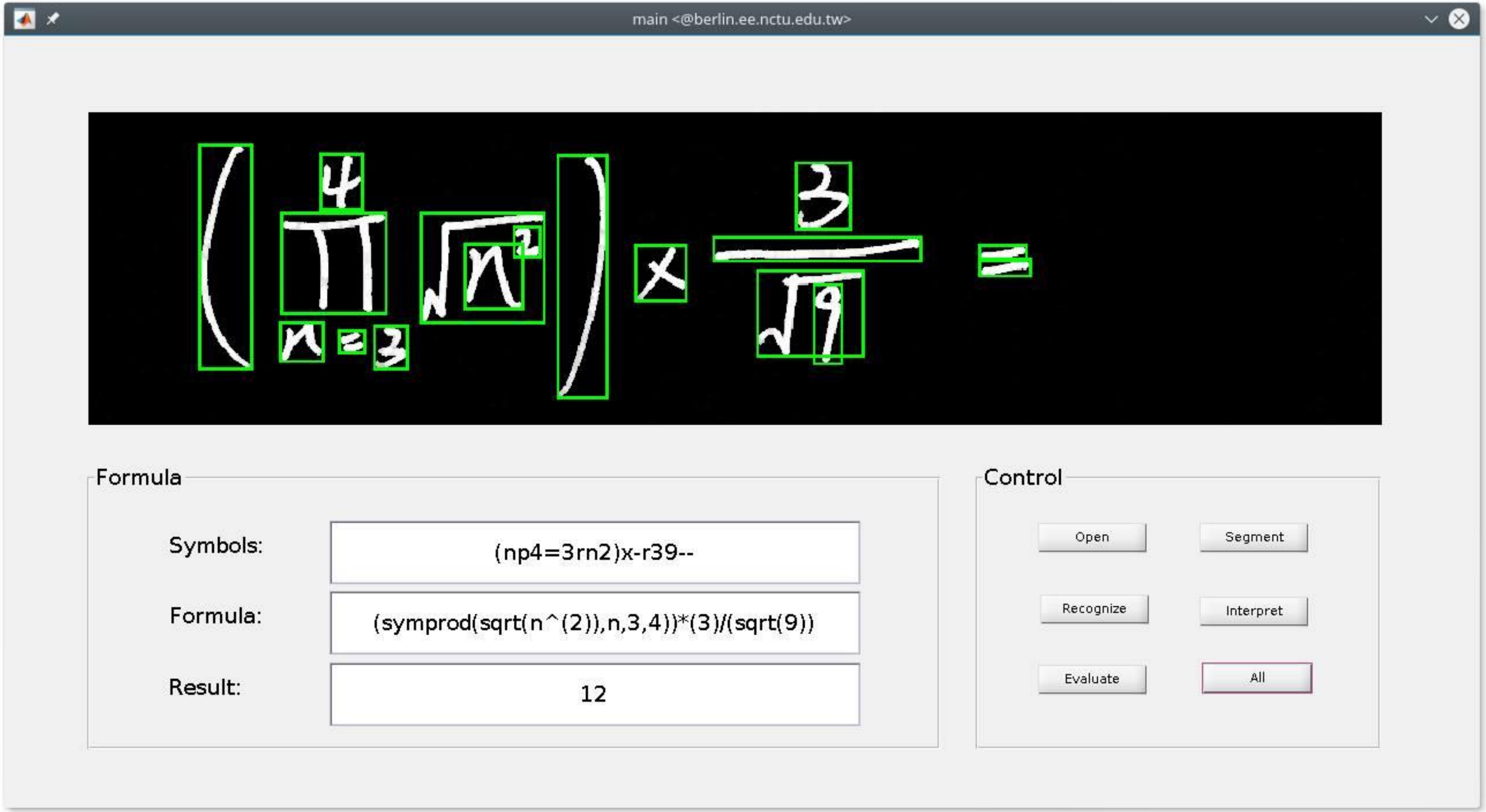
MACHINE LEARNING FINAL PROJECT
HANDWRITTEN FORMULA RECOGNITION

An Duc Pham (0450296)

Chien-Yu Lin (0450225)

Project: DICAR

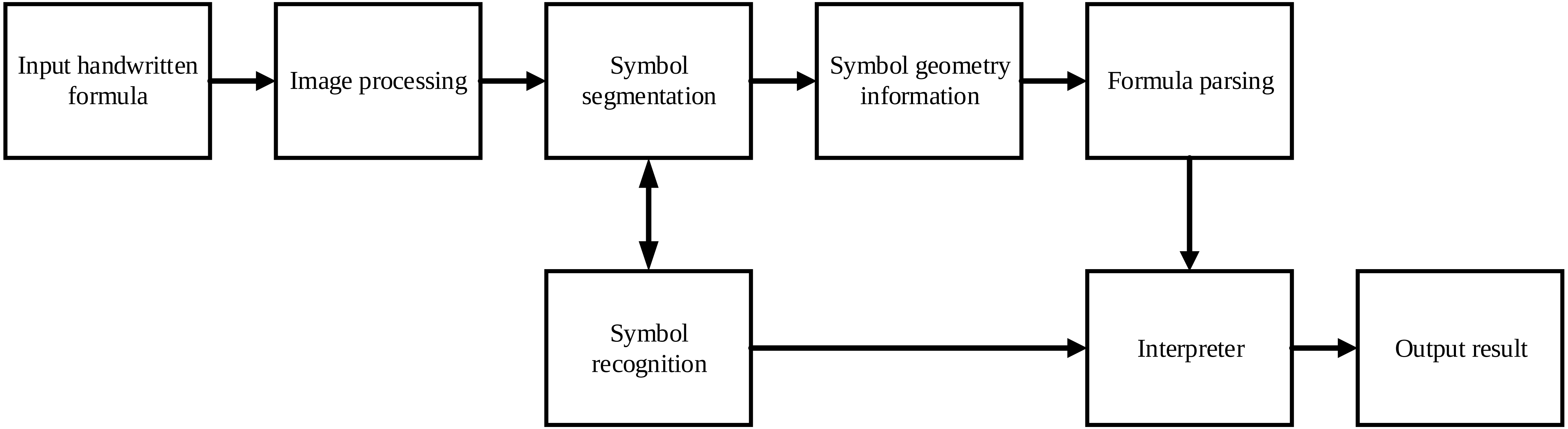


Symbols: (np4=3rn2)x-r39--

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

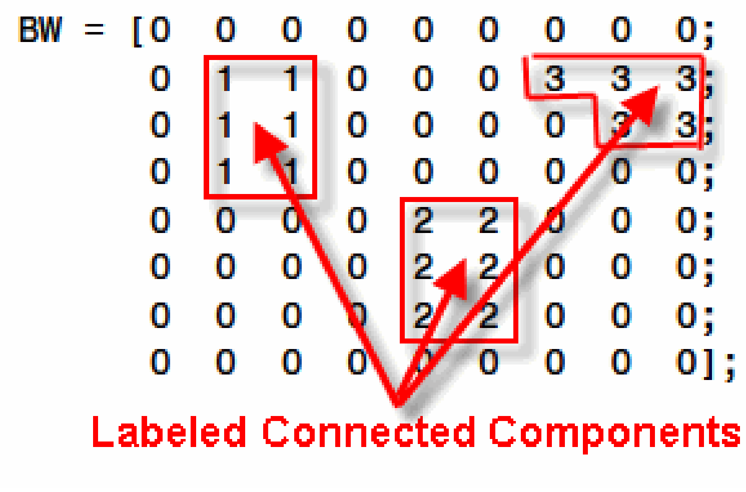
Result: 12

Proposed System



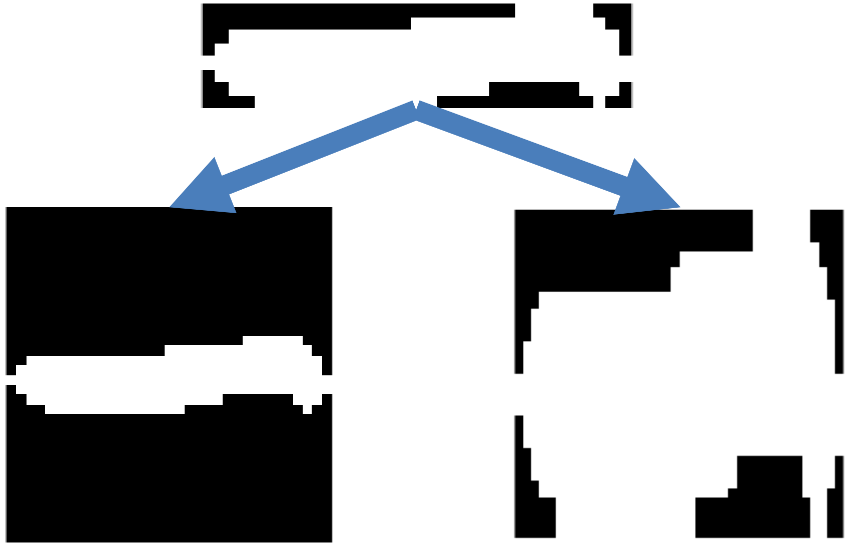
Symbol Segmentation

- Convert image to binary
- Remove noises (component less than 100 pixels)
- Segment symbol using Connected Component Labeling (CCL) algorithm
- Pad segmented symbols to improve the recognition process



```
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];
```

Labeled Connected Components

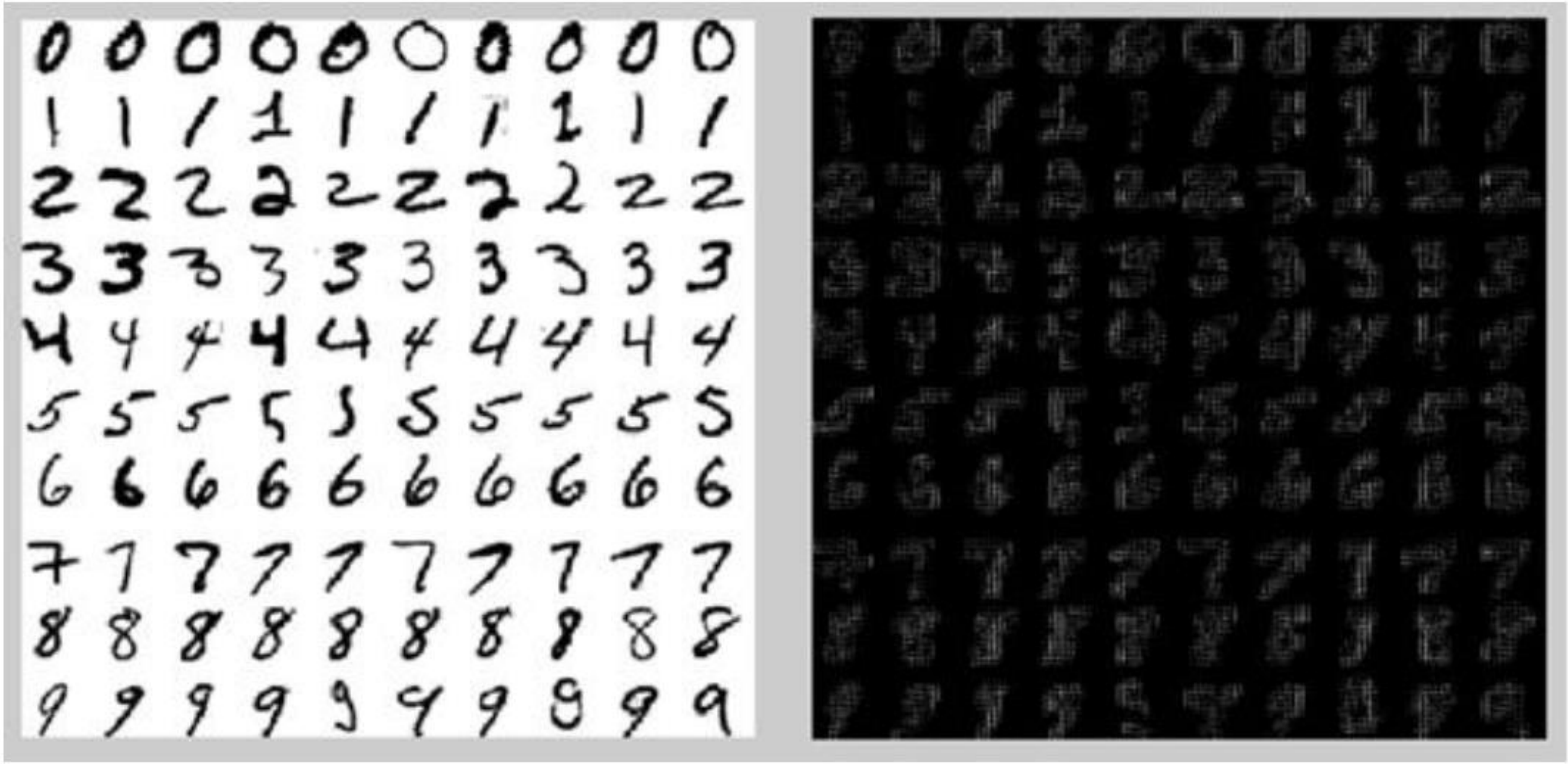


Symbol Recognition

- 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)
- 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 models which has highest probability

Symbol Recognition

- HOG for digits



Symbol Recognition

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

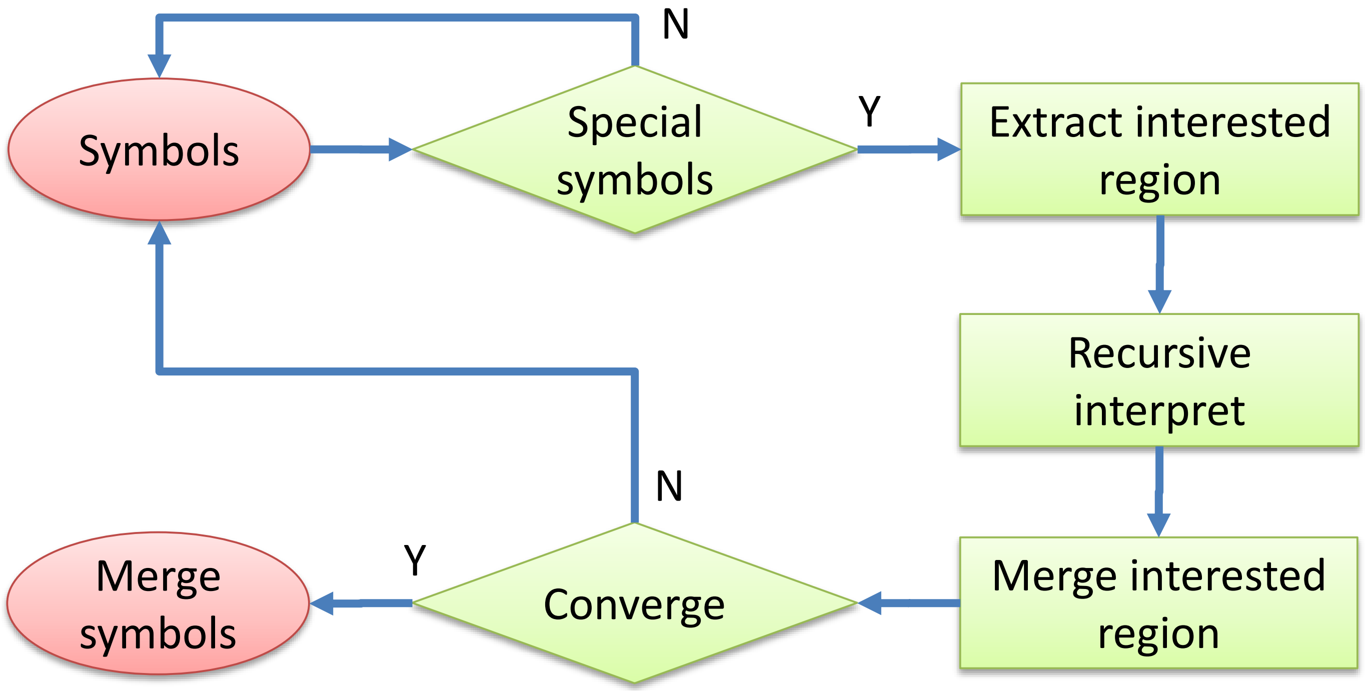
v-SVM training table

Formula Parsing

- 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: $\prod_{\boxdot}^{\boxdot}\boxdot$

Formula Parsing

- Use recursive function to update symbol data
- Repeat until the data converge



Interpreter and Evaluation

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

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

Proposed Solutions

- Use Histogram of Oriented Gradients (HOG) and PCA to extract features. Handwritten characters are based on strokes, so it could improve the recognition process
- 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