Classification of Handwritten Math Symbols using Random Forest and Hybrid Features

Kedarnath Calangutkar, Prof. Richard Zanibbi
Department of Computer Science, Rochester Institute of Technology

Introduction

- ❑ Simplify math expression input
- ❑ Draw expressions using mouse/touch to create queries
- ❑ Intuitive and almost no learning curve as compared to LaTeX, Microsoft Equation Editor, etc.

Dataset

- ❑ CROHME 2014 dataset [1]
- ❑ 101 symbol classes
- ❑ Training Data: 8836 files, 85 781 symbols
- ❑ Test Data: 986 files, 10019 symbols

Features

Fig 1. Preprocessing: Normalization, Resampling, Interpolation

- ❑ 139 features based on time series information [2]
  - 9 per point * 15 points + 4 (aggregate features)
- ❑ 102 shape based features [3]
  - Sliding window for histogram of orientations
- ❑ All the features are scaled to the range (0, 1)

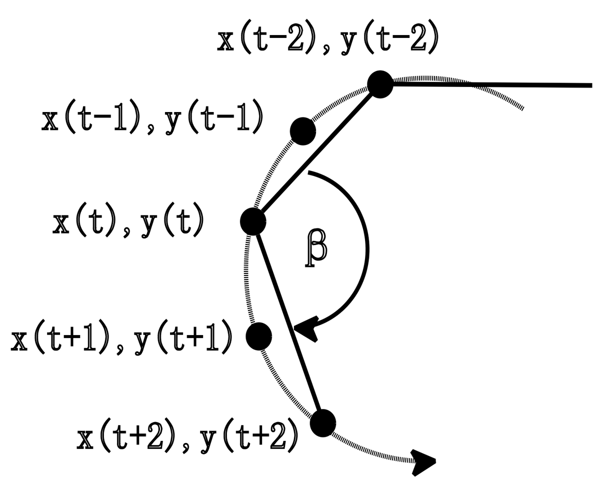

Fig 2. Angle of Curvature [2]

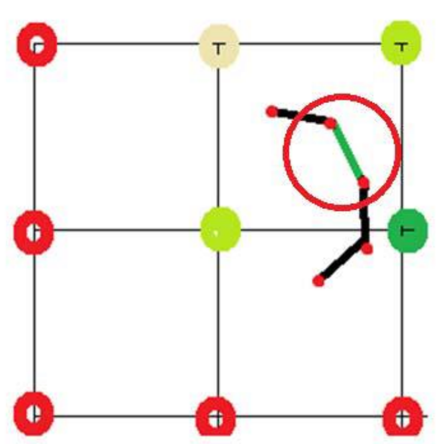

Fig 3. Histogram of Orientations [3]

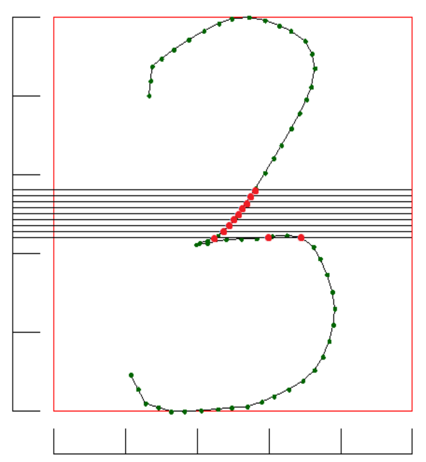

Fig 4. Histogram of 2D Crossings [3]

Classifier

- ❑ Random Forest
- ❑ Scikit-learn's implementation
- ❑ Parameters:
  - ❑ Number of decision trees in the random forest
  - ❑ Maximum depth of each decision tree

Sample expression

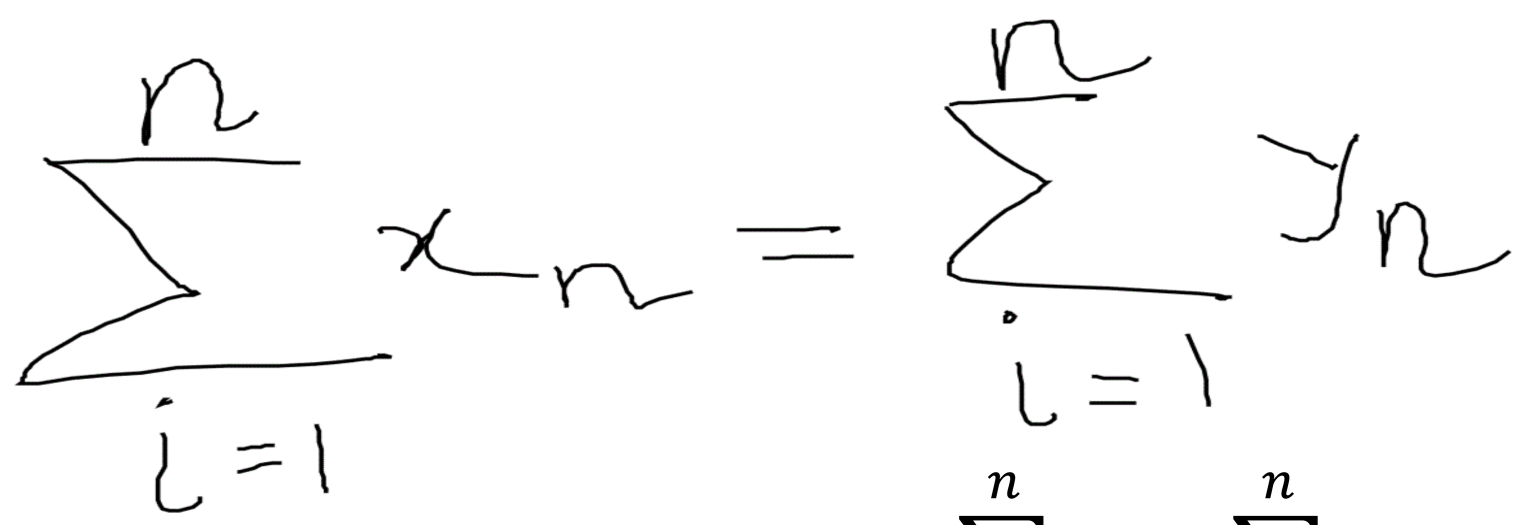

Fig 5. Expression with ground truth: $\sum_{i=1}^{n} x_n = \sum_{i=1}^{n} y_n$

Experiments

After testing different parameter values for the random forest classifier, the best training accuracy was obtained using 200 trees and a maximum depth of 18 for each tree.

By testing different resolutions of trace points for online features, it was observed that 15 trace points are enough to get high accuracy and still keep the size of the feature vector small.

Using only features based on time series information gives a training accuracy of 98.69%, while using only shape based features gives 98.36%. The combination of both gives a training accuracy of 98.72%.

Results

Table 1. Comparison with state-of-the-art methods

Recognition Systems	Recognition Rate
Universitat Politècnica de València [1]	**91.24**
Myscript (Vision Objects) [1]	91.04
DPRL, Rochester Institute of Technology [1]	88.66
Proposed SVM with linear kernel	88.15
Proposed SVM with rbf kernel	87.04
Proposed Random Forest (Hybrid features)	**88.90**
Proposed Random Forest (Online only)	87.89
Proposed Random Forest (Offline only)	87.15

Table 2. Top confusions for proposed Random Forest

Confusing Pair	No. of errors	% of total error
x, \times	174	15.64748
1, |	68	6.11510
x, X	58	5.21582
t, +	38	3.41726
1, COMMA	34	3.05755
c, C	30	2.69784
p, P	28	2.51798
(, 1	26	2.33812

Conclusion

- ❑ Combination of online and offline features produce better result than either one of those.
- ❑ Visually similar symbols are difficult to distinguish without context information.
- ❑ Using more trace points for online symbols does not improve performance after a certain point.

References

[1] Mouchère, H., Viard-Gaudin, C., Zanibbi, R., & Garain, U. (2014). ICFHR 2014 - CROHME 2014. Proceedings - IWFHR, (Crohme)

[2] L. Hu and R. Zanibbi, "HMM-Based Recognition of Online Handwritten Mathematical Symbols Using Segmental KMeans Initialization and a Modified Pen-Up/Down Feature," ICDAR, 2011.

[3] K. Davila, S. Ludi, and R. Zanibbi, "Using Off-Line Features and Synthetic Data for On-Line Handwritten Math Symbol Recognition," 2014 14th ICFHR, 2014.