Math Formula Extraction from Scholarly Documents

Ritvik Joshi (rj9936@rit.edu) Advisor: Dr. Richard Zanibbi

R·I·T — Computer Science @ RIT, B. Thomas Golisano College of Computing & Information Sciences — dprl

Goal

Develop a technique to segment character symbols and ,isolated and embedded mathematical expression from PDF document with low computational cost and with state-of the art results.

❖ Isolated Expression ❖ Embedded Expression

let angle θ between w and horizontal $\mathbf{h} = (1, 0)$ be

$$\theta = \begin{cases} \arccos\left(\frac{\mathbf{w}\cdot\mathbf{h}}{||\mathbf{w}||\,||\mathbf{h}||}\right) & \text{if } y_h \geq y_0 \\ 2\pi - \arccos\left(\frac{\mathbf{w}\cdot\mathbf{h}}{||\mathbf{w}||\,||\mathbf{h}||}\right) & \text{if } y_h < y_0 \end{cases}$$

let $\theta_{min} = \min(\theta_{min}, \theta)$, $\theta_{max} = \max(\theta_{max}, \theta)$

Math Scraper

- Retrieves character and spatial information from PDF to determine the true bounding box.
- Uses visual and geometrical features to segment math expression from normal text.

Determine True Bounding Box

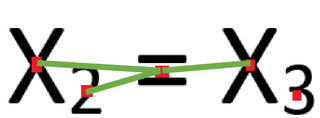
Bounding Box – Rectangular box that perfectly envelopes a character.

Word Filtering

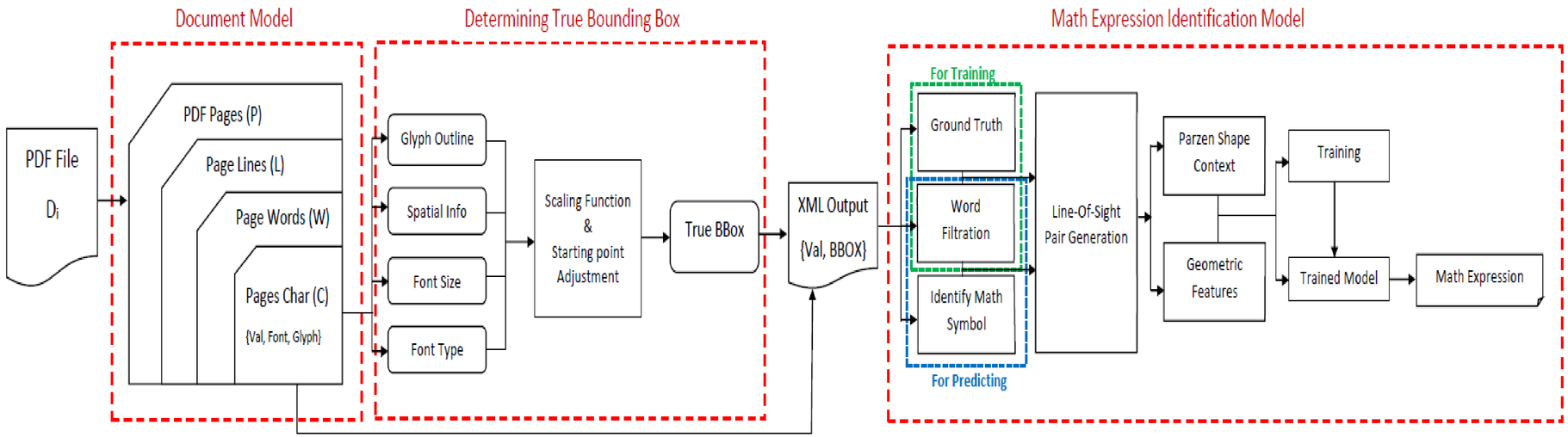
Tokenize, Stem and Flag the commonly used English words as Non-math.

Line-Of-Sight

Create pairs between the symbols which are visible from the center of current symbol.

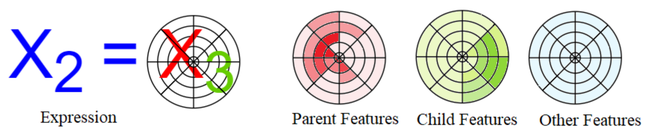


System Architecture



Parzen Shape Context

Determine pixel density between parent and child symbol using "Polar Histogram".



Results

❖ True Bounding Box

$$\inf_{x=2} \sum x + 2 \geq 10$$

PDFBOX

$$\inf_{x=2} \sum x + 2 \geq 10$$

Math Scraper

❖ Math Expression

Through a series of transformation, we will convert the above formula into what we can be estimated. Firstly, by applying Bayesian rule on $P(L|N)$, we have:

$$L_R(n) = \frac{P(N=n|L=ME)P(L=ME)}{P(N=n|L=NME)P(L=NME)} \quad (2)$$

❖ Computational Cost

Techniques	Computational Cost (Per Page)
Infty Reader	37 sec
Font Based Bayesian Model	12 sec
Math Scraper	1.9 sec

Conclusion

- Directly retrieving label information from the PDF document eliminates the need of symbol classification, thus increasing symbol classification accuracy.
- Low computational cost as it does not require any kind of image processing.

References

- L.Hu, R.Zanibbi, "Line of Sight Graphs and Parzen Shape Context Features for Handwritten Math Formula", Intl Conf. Frontiers Handwriting, 2016
- X. Lin, L. Gao, Z. Tang, and X. Hu. Mathematical formula identification in pdf documents, ICDAR 09 2011.
- Prof. Christopher Bondy – School of Print Media, R.I.T