Report-segmentation: (segmentor.py)

Suppose we are testing following image:

Before applying the model on the image, we need to cut image into pieces of symbol images:

This way, we can recognize every single symbol.

To achieve this, first we call *findCountour()* in *openCV* class. This gives us a list of contours. Using this list, we cut image pieces from the image. Now we have:

Here are two problems:

1. Symbols having separate parts won't be recognized, like "i" and "=".
2. For single image piece that is cut out, we may have something we don't want, like:

while we need "sqrt" only.

To solve problem 2: we used label() to convert each continue shape into single digits. Since there is only 1 such digit appearing on all four edges of the image piece, we can find it and keep it only. This way, we can erase "2" and have:

To solve problem 1: we check if the image piece is bar-shaped. If so, we conjugate it with other bar pieces. If they mostly overlap horizontally, we combine them. This way, we have:

 and 

The code for this part is relatively complicated since we need to extract coordinates of

bars constantly and at the same time keep bars list.

Such segmentation method can recognize most of the “=” and some of the “i”, while only does few mistakes (e.g. long sqrt and long frac inside). However, combining “division sign” is still a hard task for current code because the dots are too far from the bar and it would be inefficient to check this case for all possible conjugations.