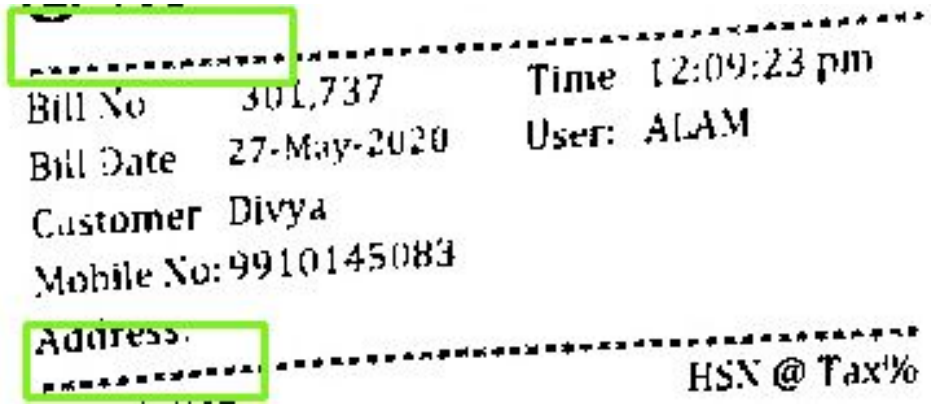
INVOICE OCR

PROCEDURE

1. Pre-Processing steps:
    i. Grayscaling
    ii. Binarisation
    iii. Foreground as white and background as black
    iv. Blurring
    v. Dilation for effective contour detection
    vi. Line Detection using Hough Line method (detecting boundary of receipt)
    vii. Perspective Transformation
    viii. Again, trying to find contours on the transformed image which will give better results as compared to the previous run.
    ix. Extract client name here only (assuming the name of the store will bein 3rd contour always, though it can be generalized further)
    x. Zooming and cropping to get this region because my assumption is that I will be able to find my interest text between the two " ********** " star lines.



xi. So, sorting top 5 bounding boxes/contours on the basis of width, so that first and second most are the 2-star regions.

xii. Finally cropping to this image After detecting contours, I can feed this region to tesseract by the fine-tuning configuration of tesseract and I will be able to get the required text.

xiii. Tesseract text output was :

" Bill o OO Th
iil Date 27-May-2020 Us
Customer Divya
Mobile Nor99 10145083

xiv. Now, I need to apply Regular Expression patterns to extract fields.