С помощью различных источников в интернете мы смогли, используя библиотеку OpenCV, разбить фото таблицы на ячейки. Для анализа содержимого ячеек мы используем библиотеку pytesseract, а именно функцию image\_to\_string(). Ну и дальше вносим данные в соответствующие ячейки excel.

Но пока эффективность оставляет желать лучшего. Тщательной доработки мы ещё не проводили, только на последних этапах выявились очевидные проблемы.

Первой проблемой является разметка таблицы: лишние прямые пока включаются в линии таблицы, что портит исход. В этой таблице например:

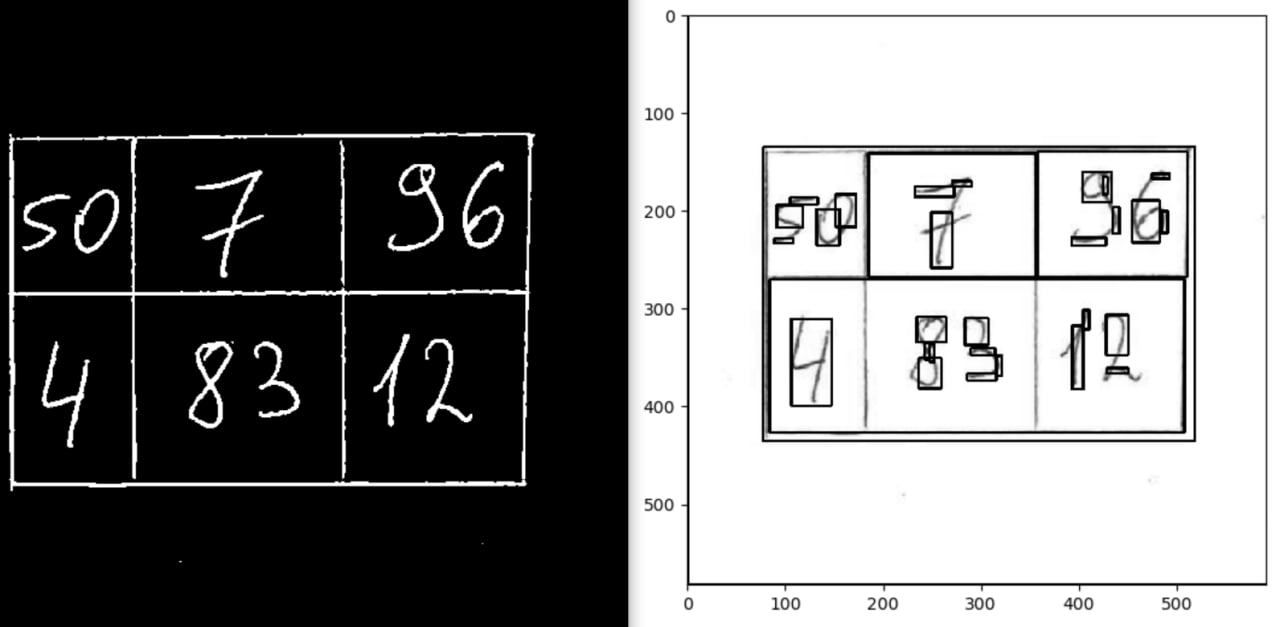
Изображение выглядит как текст, квитанция

Автоматически созданное описание

Изображение выглядит как текст, здание

Автоматически созданное описание

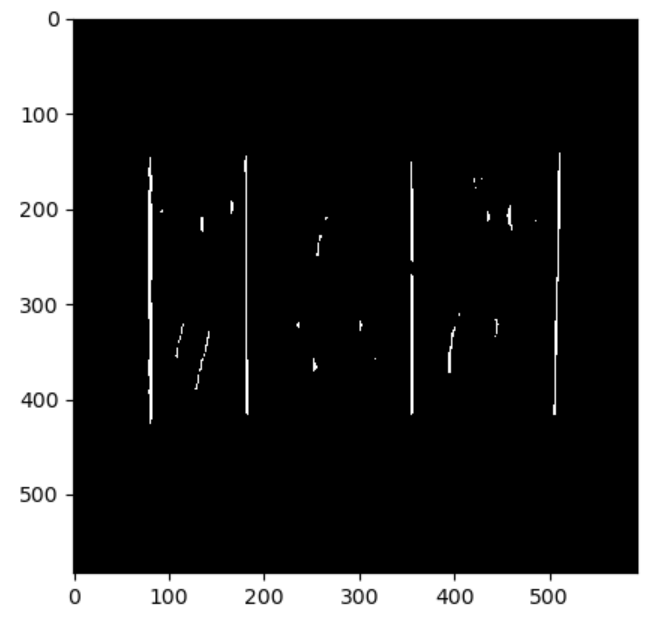
Также мы создали свои простые таблицы для тестов, но теперь проблема в том, что функции, находящие горизонтальные и вертикальные линии, распознают эти прямые в цифрах, и соответственно разметка таблицы портится. Например, в таких:



Второй большой проблемой является некорректное распознавание содержимого ячеек: функция image\_to\_string(). Она даже не видит порой ничего – ячейки в excel остаются пустыми.

Можете подсказать что-то по нашим проблемам. Есть ли какой-то алгоритм, для избавления от ненужных линий?

Почему вообще OpenCV видит горизонтали и вертикали в цифрах на второй таблице, а на первой такого нет? Программа вот такие вертикали находит:



И есть ли способ улучшить читаемость содержимого ячеек? Почему pytesseract так плохо работает, ведь мы используем модуль Tesseract-OCR от гугла?