Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего профессионального образования

**"Национальный исследовательский
университет Высшая школа экономики"**

# Распознавание протоколов

Отчет по проекту

Профориентационный семинар

Введение в специальность

Выполнили Кармаев А.А. Родин С.А. Индюченко Н.А.

Принял Сластников Сергей Александрович

Москва 2022

Содержание:

## Введение

В настоящее время все больше используются электронные системы хранения данных. Но многие организации все еще используют консервативные методы хранения информации на бумажных носителях. Благодаря современным технологиям стало возможно конвертировать печатные и рукописные документы в электронный формат. Обеспечение системами распознавания табличных данных позволяет значительно оптимизировать процесс работы и избавить сотрудников от ручного переноса данных с бумажных носителей на электронные..

## Цель работы

Разработка и тестирование программы по переводу рукописно заполненных таблиц с протоколами различных соревнований в excel таблицу.

## Предмет исследования

Различные инструменты распознавания образов и детектирования содержимого определенного вида печатных и рукописных документов на основе технологий Computer Vision.

## Задачи

- Изучить технологии распознавания объектов, текста для обработки документов, выбрать наиболее подходящие для нашей цели инструменты;
- Разработать алгоритмы обнаружения и распознавания таблицы в протоколе, используя выбранные выше технологии;
- Протестировать полученные решения на разных типах входных данных для получения метрик точности распознавания

## Планируемые результаты

- Реализация различных алгоритмов решения задачи и выявление наиболее эффективного способа считывания документов.
- Сравнение результатов работы программы с печатными и с рукописными данными.
- Определение наиболее подходящих параметров исходных изображений и настроек программы для улучшения распознавания

# Теория

## OCR

OCR — это технология, которая распознает текст в цифровом изображении. Она обычно используется для распознавания текста в отсканированных документах, но также служит многим другим целям.

Программное обеспечение OCR обрабатывает цифровое изображение, обнаруживая и распознавая символы, такие как буквы, цифры и символы. Некоторые программы распознавания текста просто экспортируют текст, в то время как другие программы могут преобразовывать символы в редактируемый текст прямо на изображении. Расширенное программное обеспечение OCR может экспортировать размер и формат текста, а также макет текста, найденного на странице.

Технология OCR может использоваться для преобразования печатной копии документа в электронную версию. Например, если вы сканируете многостраничный документ в цифровое изображение, такое как файл TIFF, вы можете загрузить документ в OCR программу, которая распознает текст и преобразует документ в редактируемый текстовый файл.

Так как для реализации алгоритма распознавания информации с оценочной ведомости был выбран язык программирования Python, то были рассмотрены библиотеки OCR, совместимые с ним – PyTesseract.

Для данного проекта был выбран для использования Python-Tesseract или просто PyTesseract, библиотека, которая является оболочкой для Google Tesseract-OCR Engine.

## Нейронные сети.

Основная идея нейронной сети состоит в том, чтобы моделировать (копировать в упрощенном, но достаточно точном виде) множество плотно связанных между собой клеток мозга внутри компьютера, чтобы иметь возможность научить его распознавать вещи, закономерности и принимать решения по-человечески. Удивительная вещь о нейронной сети состоит в том, что ее не нужно программировать для явного обучения: она учится сама, как человеческий мозг.

Для задач распознавания можно выделить два основных типа нейронных сетей – сверточные (Convolutional neural network) и рекуррентные (Recurrent neural network). Первые идеально подходят для обработки изображений и видео, вторые для анализа текста и речи. В рамках проекта были рассмотрены сверточные нейронные сети, так как для выполнения задач потребуется работа с изображениями табличных протоколов.

## TensorFlow

Созданная Google и написанная на Python и C++, TensorFlow является одной из лучших открытых библиотек для численных вычислений. Ее качество признано на рынке, поскольку даже такие гиганты как DeepMind, Uber, AirBnB или Dropbox выбрали этот фреймворк для своих нужд. TensorFlow хороша для сложных проектов, таких как создание многослойных нейронных сетей. Она используется для распознавания голоса или картинок и приложений для работы с текстом, таких как Google Translate, например.

Преимущества TensorFlow:

- Большое количество руководств и документации;
- Предлагает мощные средства мониторинга процесса обучения моделей и визуализации (TensorBoard);
- Поддерживается большим сообществом разработчиков и техническими компаниями;
- Обеспечивает обслуживание моделей;
- Поддерживает распределенное обучение;
- Использование различных мощностей компьютера (GPU, поддерживающие технологию CUDA, CPU и TPU)

Недостатки:

- Имеет более высокий входной порог для начинающих, чем Keras. Голая Tensorflow достаточно низкоуровневая и требует много шаблонного кода, и режим «определить и запустить» для Tensorflow значительно усложняет процесс дебага.
- Единственный полностью поддерживаемый язык – Python.

## CUDA

CUDA - программно-аппаратная архитектура параллельных вычислений, которая позволяет существенно увеличить вычислительную производительность благодаря использованию графических процессоров фирмы Nvidia. Функции, ускоренные при помощи CUDA, можно вызывать из различных языков, в том числе Python.

Для полной поддержки технологии CUDA, которую мы хотим использовать для обучения нашей модели, необходимо:

1) Наличие видеокарты Nvidia, которая поддерживает CUDA. В нашем случае – это видеокарта GTX1650.

2) Поколение видеокарты должно быть не старее Kepler – у нас Pascal.

3) Обновить драйвер до последней версии (выше 410.0). В нашем случае установлена версия 516.01.

4) В зависимости от версии драйвера установить соответствующую версию CUDA. Для нас актуальна версия 11.7.

5) Установить соответствующую версию CUDA Toolkit, которая связана с CUDA и версией драйвера вашей видеокарты.

6) Установить и настроить библиотеку cuDNN, которая также связана с версией CUDA.

**Примечание:** Версия Tensorflow-GPU должна быть 2.0 или выше для поддержки мощностей видеокарты Nvidia.

## TPU и Colaboratory

**TPU** - тензорный процессор, относящийся к классу нейронных процессоров, являющийся специализированной интегральной схемой, разработанной корпорацией Google и предназначенной для использования с библиотекой машинного обучения TansorFlow. Тензорный процессор состоит из четырех чипов, каждый из которых содержит два ядра, всего в TPU восемь ядер. Обучение на TPU ведется параллельно на всех ядрах с помощью репликации: на каждом ядре работает копия графа TensorFlow с одной восьмой объема данных.

**Colaboratory** - это облачная платформа от Google для продвижения технологий машинного обучения. На ней можно получить бесплатно виртуальную машину с установленными популярными библиотеками TensorFlow, Keras, sklearn, pandas и т.п. Самое удобное, что на Colaboratory можно запускать ноутбуки, похожие на Jupyter. Ноутбуки сохраняются на Google Drive, можно их распространять и даже организовать совместную работу.

Именно поэтому в командной работе мы воспользуемся Colaboratory, так как помимо удобности в реализации проекта, большим плюсом являются мощности серверов. Наша система позволяет воспользоваться всего лишь 6 TFlops. То есть облачный сервис способен в 30 раз быстрее обучить нашу нейронную сеть, чем настольный компьютер. Однако есть и свои минусы:

1) Данные нужно хранить на Google Drive, так как виртуальная машина предоставляется на протяжении 12 часов, после чего все данные удаляются. Однако ничто не мешает запустить ещё одну виртуальную машину.

2) Потери производительности могут быть связаны из-за низкой скорости интернета, которая связывает ПК и сервером, то есть является bottleneck’ом (узким местом нашей цепочки).

## Keras

Keras – это минималистичная библиотека, основанная на Python, которая может запускаться поверх TensoFlow, Theano или CNTK. Она была разработана инженером компании Google, Франсуа Шолле, в целях ускорения экспериментов. Keras поддерживает широкий спектр слоев нейронных сетей, таких как сверточные слои, рекуррентные или плотные. Этот фреймворк хорош в кейсах для перевода, распознавании изображений, речи и т. п.

Преимущества:

§ Быстрое и простое прототипирование;

§ Достаточно маловесный для построения моделей глубокого обучения для множества слоев;

§ Наличие полностью конфигурируемые модули;

§ Простой и интуитивно-понятный интерфейс, соответственно, хорош для новичков;

§ Встроенная поддержка для обучения на нескольких GPU;

§ Поддержка GPU от NVIDIA, TPU от Google, GPU с Open-CL, таких как AMD.

Недостатки:

§ Может оказаться слишком высокоуровневым и не всегда легко кастомизируется;

§ Ограничен бэкэндами Tensorflow, CNTK и Theano

В рамках данного проекта для обучения сверточной нейросети, предназначенной для распознавания рукописных цифр, был выбран модуль для глубокого обучения Keras, так как он является легким, функциональным и поддерживает часть возможностей библиотеки TensorFlow.

## Технологии работы с изображением

### 1. PIL (Python Imaging Library)

Python Imaging Library расширяет возможности обработки изображения для языка программирования Python. Данная библиотека обеспечивает обширную поддержку различных форматов файлов, довольно мощные возможности обработки изображений и должна обеспечить прочную основу для общей обработки изображений.

PIL содержит некоторые основные функции предобработки изображений, такие как фильтрация с набором встроенных ядер свертки и преобразование цветового пространства. Библиотека также поддерживает изменение размера изображения, поворот и произвольные аффинные преобразования.

### 2. OpenCV

OpenCV зародился в корпорации Intel в 1999 году благодаря Гэри Брэдски, а первый релиз состоялся в 2000 году. Сейчас, OpenCV обладает обширными возможностями, связанными с компьютерным зрением и машинным обучением, и расширяется с каждым днем.

OpenCV поддерживает широкий спектр языков программирования, таких как C++, Python, Java и др., и доступен на различных платформах, включая Windows, 22 Linux, OS X, Android и iOS. Активно разрабатываются интерфейсы для высокоскоростных операций GPU на основе CUDA и OpenCL.

OpenCV-Python - это API Python для OpenCV, сочетающий в себе лучшие качества API OpenCV C++ и языка Python.

OpenCV-Python использует библиотеку Numpy, которая является высоко оптимизированной библиотекой для числовых операций с синтаксисом в стиле MATLAB. Все массивы OpenCV преобразуются в массивы Numpy и наоборот. Это также упрощает интеграцию с другими библиотеками, использующими Numpy, такими как SciPy и Matplotlib.

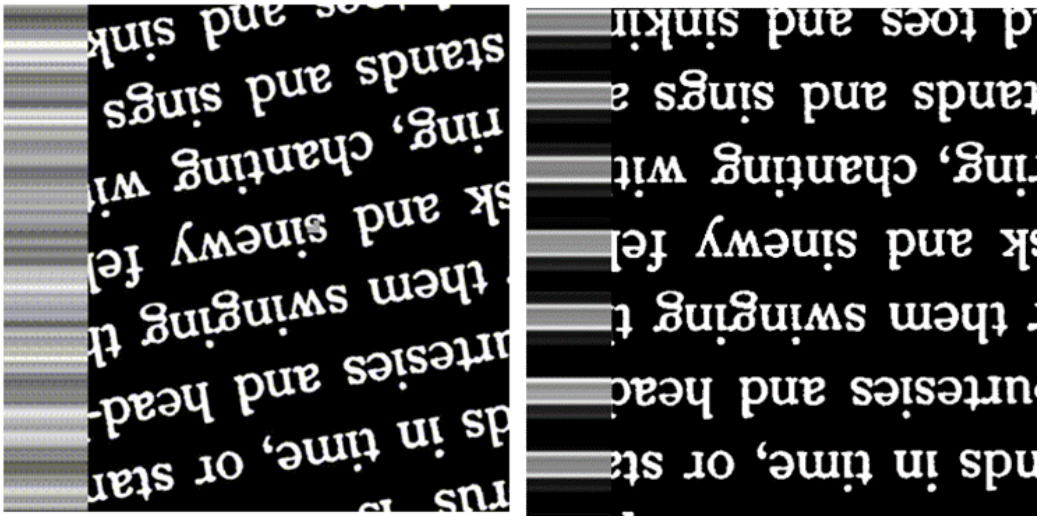
В данной библиотеке содержатся возможности, аналогичные возможностям PIL, но, кроме этого, она наделена гораздо большими возможностями, такими как компьютерное зрение и глубокое обучение.

# Практика.

## Поворот картинки.

Первая задача, которую нужно решить при обработке фотографии – это выявление и устранение наклона таблицы. Для этого используется метод суммирования строк.

Фотография – это двумерный массив, хранящий описание пикселей. При решении проблемы поворота мы работаем с уже преобразованным к бинарной форме изображением, следовательно, каждый пиксель может принимать значение 0 или 1. Мы можем суммировать значения в каждой строке и получить одномерный массив со значениями “насыщенности цвета” для каждой строки. Данный массив в коде программы сохраняется в переменную histogram



Далее, нужно определить дельту между максимальным и минимальным значениями в массиве. Чтобы определить угол, на который повернута таблица, перебираем углы от 0 до 360 с определенным шагом и запоминаем угол с максимальной дельтой. Это и будет угол, на который повернуто наше изображение.

```
score = np.sum((histogram[1:] - histogram[:-1]) ** 2, dtype=float)

best_angle = angles[scores.index(max(scores))]
```

Для поворота изображения в библиотеке OpenCV существуют функции getRotationMatrix(center, angle, scale) и warpAffine(img, M, size, …).

```
M = getRotationMatrix(center, angle, scale)
```

center – центр изображения

angle – угол, на который изображение должно быть повернуто

scale – множитель увеличения/уменьшения размера изображения

M – матрица преобразования, применяемая к изображению

img = warpAffine(img_init, M, size, …)

img_init – исходное изображение

M – матрица преобразования

size – размер полученного изображения

img – преобразованние изображение

С помощью этих функций:

1. Находим матрицу преобразования
2. Считаем размер изображения после поворота, чтобы избежать обрезания краев:
3. Применяем афинное преобразование к фотографии, чтобы развернуть ее в правильное положение:

Таким образом, фотография готова для дальнейшего считывания таблицы.

## Использование OpenCV и PyTesseract для считывания таблицы.

Алгоритм решения задачи делится на два ключевых этапа: 1) разбиение таблицы на ячейки и 2) распознавание содержимого каждой. И первый пункт является самым объёмным, из-за универсального подхода к распознанию структуры таблицы и нужды предусматривать различные исключения на каждом этапе.

Основные функции применяемые в программе (main.py) хранятся в модуле package.py

Первым делом считываем фото и производим предобработку исходной картинки для более точного распознавания -> reading_photo(…):

1. Считываем фото из файла -> cv2.imread(…)

2. Поворот картинки с помощью функции rotate_to_correct(…) по алгоритму описанному выше
3. Увеличивается размер фото минимум до 600 пикселей в высоту -> cv2.resize(…)
4. Сводим фото к оттенкам серого
5. Размываем фото -> cv2.erode(…)
6. Избавляемся от шумов, сведением фото к двум цветам - чёрному(фон) и белому(текст) -> cv2.adaptiveThreshold(…)

**Разбиение таблицы на ячейки:**

1. Распознание вертикальных линий, формирующих таблицу -> vertical_lines_detection(…)

Основой для выполнения конкретной задачи является функция cv2.getStructuringElement(cv2.MORPH_RECT, (1, kernel_len)), которая распознаёт на фото прямоугольники с размером ядра, задающим вертикальную линию - минимальная ширина по оси X

2. Распознание горизонтальных линий, формирующих таблицу -> horizontal_lines_detection(...)

Данная часть выполняется аналогично функцией cv2.getStructuringElement(cv2.MORPH_RECT, (kernel_len, 1)), которая распознаёт на фото прямоугольники с размером ядра, задающим горизонтальную линию - минимальная ширина по оси У

3. Объединение горизонталей и вертикалей, для формирования цельной таблицы без содержимого -> combining_horizontals_and_verticals(...)

Применяем функцию cv2.addWeighted(vertical_lines, 0.5, horizontal_lines, 0.5, 0.0), которая наложит полученные ранее фото вертикальных и горизонтальных линий в равном соотношении прозрачности

4. Находим контуры таблицы, полученной на предыдущем шаге, таким образом мы получим контуры всех прямоугольников таблицы - наши будующие ячейки -> cv2.findContours(...)

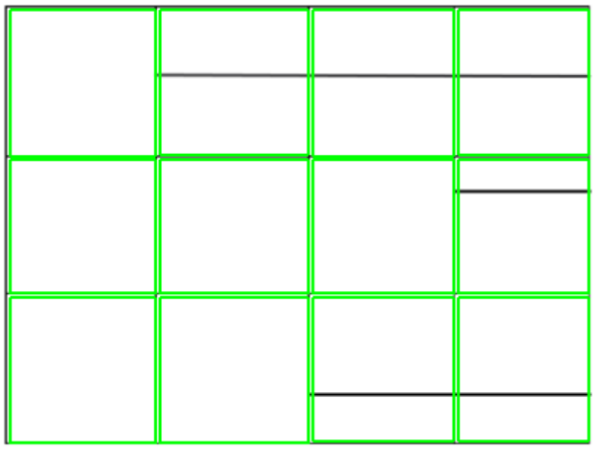
Используем режим группировки CV_RETR_TREE — группирует контуры в многоуровневую иерархию и метод упаковки CV_CHAIN_APPROX_SIMPLE — склеивает все горизонтальные, вертикальные и диагональные контуры.

5. Сортируем контуры таблицы по порядку их расположения в таблице (сверху вниз) -> sort_contours(...)

6. Формируем из контуров список ячеек с рамками -> form_cells(...)

Данный этап программы можно разделить на подпункты, ведь именно здесь происходит обработка всевозможных исключений, который позволят исключить из списка контуров ненужные - те, которые были выявлены или из-за ошибки программы, или из-за заранее известного фактора

   a. Имея список контуров переводим их в список прямоугольников задающих координаты ячеек таблицы -> cv2.boundingRect(...), и добавляем в список лишь те, значения ширины и высоты которых не превосходят половины от размеров всего фото - такие ячейки будут представлять из себя контур всей таблицы, а она не является ячейкой таблицы, которую стоит рассматривать
   b. Удаляем ячейки, которые лежат внутри других - они определенно лишние, такого не должно быть в таблице. Подобное может возникнуть из-за большого разрешения содержимого ячеек, которые могут сами быть приняты за таковые, так как в них могут распознаваться вертикали и горизонтали
   c. Проверяем лежат ли две ячейки в одном столбце и в одной строке - погрешность, из-за черты, которую человек мог провести в какой-то строке. И мы удаляем нижнюю из этих двух, а верхнюю ячейку расширяем, добавляя высоту нижней



7. Разбиваем список рамок на ячейки по строкам -> rows(...)

До тех пор, пока поле не отличается больше, чем его собственное (высота + среднее значение / 2), поле находится в том же ряду. Как только разница в высоте становится больше текущей (высота + среднее значение /2) мы знаем, что начинается новая строка

8. Сортируем строки в каждом столбце (слева направо) -> rows_sort(...)

**Распознавание значений в ячейках -> recognition(...)**

1. Рассматриваем циклично содержимое каждой ячейки
2. Подгоняем содержимое для лучшего распознавания расширением и размыванием -> cv2.dilate(...) & cv2.erode(...)
3. Анализируем содержимое с помощью технологий компьютерного зрения и формируем список значений на соответствующих позициях -> pytesseract.image_to_string(...)

Язык для распознавания содержимого устанавливаем русскаий: lang='rus'

Также делаем дополнительную обработку ячейки, если её значение пустует. Это может означать, что ячейка содержит единственный символ, который функция pytesseract.image_to_string(...) не способна распознать при базовой конфигурации. Поэтому необходимо условие с дополнительным считыванием при следующих параметрах:

```
pytesseract.image_to_string(erosion, lang='rus', config='--psm 10 --oem 1 -c tessedit_char_whitelist=123456789+-')
```

Последним пунктом, конечно же, является запись матрицы значений в excel файл -> data_to_excel(...)

**Вспомогательное программное обеспечение**

Программа написана на языке программирования Python с использованием различных его модулей:

Matplotlib.pyplot – модуль для визуализации данных: выводим распознанные горизонтали и вертикали, саму таблицу, ячейки с рамками

NumPy – необходим для упрощения работа со списками данных

Pandas – библиотека применяется для записи таблицы распознанных значений в excel документ

Scipy – модуль применяется в функции поворота картинки

## Использование TensorFlow и PyTesseract для считывания таблицы.

Данный способ считывания таблицы с фотографии представляет собой обучение и использование сверточной нейронной сети с помощью tensorflow и Keras для

определения положения и границ таблицы и применение pytesseract для непосредственного извлечения данных из нее.

Основные составляющие программы:

1. Подготовка Датасета для обучения нейронной сети.
2. Описание модули и ее обучение.
3. Применение модели для распознавания таблиц на фотографиях.

## Подготовка датасета для обучения сверточной нейронной сети.

Для обучения сверточной нейронной сети, которая должна находить таблицу на фотографии, используется датасет под названием “Marmot Dataset”. Он содержит изображения и их описания в формате xml. В каждом описании хранятся:

1. координаты и размеры таблицы
2. координаты и размеры столбцов таблицы

Пример изображения из датасета:

17 students taking an MSc in Human Centred Systems took part. 8 had taken an MSc module in User Modelling. All had Pocket PCs. 10 undergraduate students taking a degree in Computer Interactive Systems, who had completed undergraduate modules on Personalisation and Adaptive Systems and Interactive Learning Environments, voluntarily took part. Data was obtained by anonymous questionnaire from all subjects, and anonymous logbooks on Pocket PC use over 6 weeks from MSc students. Due to the low numbers it is inappropriate to perform a statistical analysis of the results: the aim is to discover if initial data indicates further work to be valuable.

### 2.1 Results

**Location-Aware User Modelling System.** Logbook data shows the most common location of Pocket PC use to be at home, followed by various rooms in EECE. Some students also used their Pocket PC in other parts of the campus and elsewhere. Results of 3 typical users are presented in Table 1, as an example of similarities and differences between Pocket PC use. 10 of the generally common activities are listed: reading, email, web browsing, notes, calendar, computer assisted learning, word processing, calculator, music, games. Each user also performed a few additional tasks in other categories, not shown (e.g. MSN Messenger, Excel, viewing lecture slides).

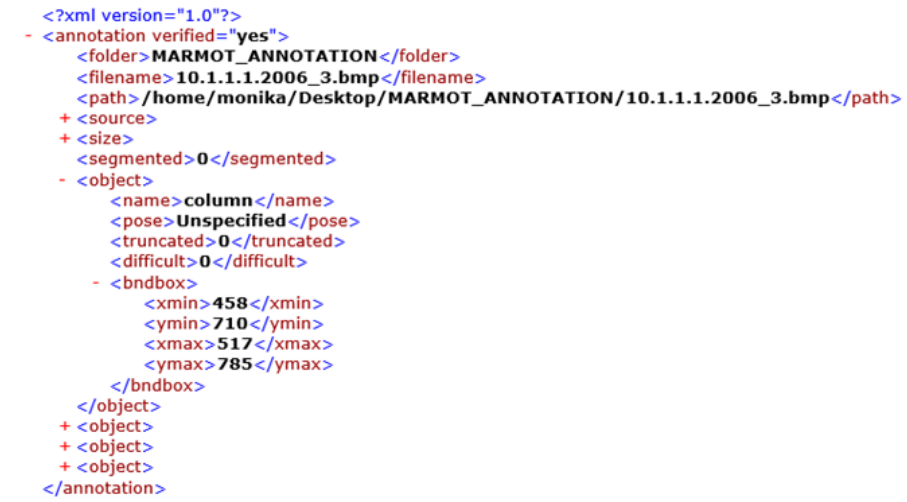
Table 1: Activities and location of use of Pocket PC by 3 students

| | *Location* | *read* | *mail* | *web* | *note* | *cal* | *CAL* | *WP* | *calc* | *mus* | *game* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| S1 | home | 1 | 2 | | 5 | 4 | 2 | 2 | 1 | 7 | 1 |
| | EECE G16 | 1 | 2 | 1 | 4 | 1 | | | 1 | | |
| | EECE 337 | | 1 | 1 | | 1 | | | 1 | 2 | |
| | EECE 421 | 1 | | 1 | | | | | | 1 | |
| | EECE 435 | | | 2 | | | | | 2 | | |
| | EECE CR | 1 | 3 | 3 | | | | | | | |
| | EECE lib | | 3 | 1 | | 1 | | | | 1 | |
| | main lib | | | 3 | 1 | | 1 | | | | |
| | shop | | | | 4 | | | 1 | | | |
| S2 | home | 1 | 4 | 3 | 1 | 4 | 2 | 2 | | 7 | 3 |
| | other home | | 1 | | | 1 | | 1 | | 1 | |
| | EECE 123 | | 1 | 1 | | | | | | | |
| | EECE 337 | | 5 | | | | | | | 1 | |
| | EECE 522 | | | | 1 | 1 | | | | | |
| | EECE CR | | 1 | | | | | | | | |
| | EECE lib | | | | | | | 1 | | | |
| | EECE rec | 1 | 2 | | | | | | | | |
| | campus | | 1 | | | 1 | | | | | |
| | restaurant | | 1 | | | 1 | | | | 1 | |
| | train | | 3 | | | | | | | 1 | 1 |
| S3 | home | 1 | 1 | | 20 | 4 | 2 | | | 4 | 8 |
| | EECE 225 | | 1 | 2 | | | | | | | |
| | EECE 337 | | | 2 | | | | | | | |
| | EECE 421 | 1 | 1 | 1 | 9 | | | | | 1 | 1 |
| | learn centre | | | 1 | 1 | | | | | | |

Пример описания таблицы:

```
<?xml version="1.0" encoding="UTF-8"?>
- <Page PageType="3" PageNum="2" CropBox="0000000000000000 408a500000000000 4082980000000000 0000000000000000">
  - <Contents>
      + <Leafs Label="Char">
      + <Leafs Label="Path">
      + <Composites Label="Textline">
      + <Composites Label="TableCaption">
      + <Composites Label="TableBody">
      - <Composites Label="Table">
          <Composite Label="Table" PLID="3804" LID="3798" BBox="405d5eb851eb851f 407c4bab4f0d844d 407ddc7ae147ae15 4064370a3d70a3d7" CLIDs="3796 3797"/>
        </Composites>
      + <Composites Label="Paragraph">
      + <Composites Label="Body">
    </Contents>
</Page>
```

Пример описания столбцов таблицы:




На аннотациях к фотографиям с таблицами содержится информация о расположении самой таблицы и о расположении ее столбцов.

В аннотациях расположение таблицы закодировано шестнадцатеричным числом, поэтому нам нужно сначала его декодировать с помощью функции hex2double.

def hex2double(hexcode):

hexcode – шестнадцатеричное число

return координаты расположения таблицы

Для определения границ таблицы по описанию из датасета используется функция return_bboxes_v1.

def return_bboxes_v1(xml, width, height):

xml – путь к файлу в формате xml

width, height – размеры изображения

return bboxes – координаты границ таблицы

Аналогично выглядит функция определения границ столбцов таблицы.

def return_bboxes_extended(xml):

xml - путь к файлу в формате xml

return width, height – размеры изображения

bboxes – координаты границ таблицы

Далее, с помощью функции create_mask создается маска таблицы, на которой показано ее расположение на фотографии, а также маска столбцов.

def create_mask(img_path, list_of_bboxes, dim, mode='table'):

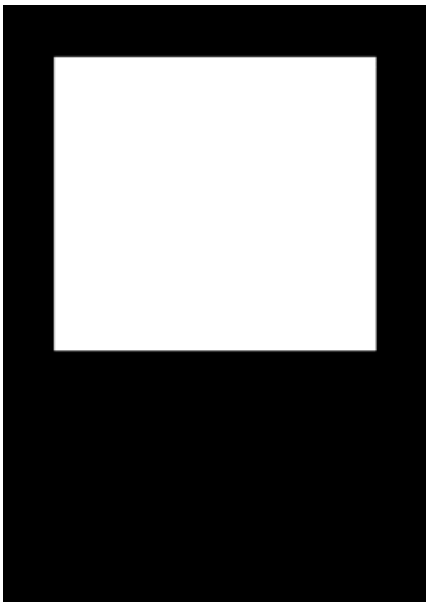
img_path – путь к изображению

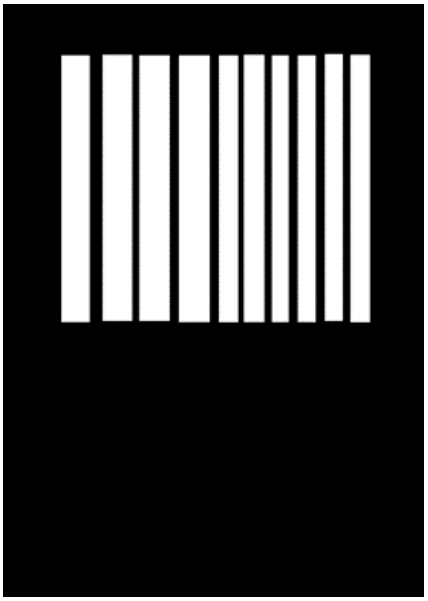
list_of_bboxes – список, хранящий границы таблицы и столбцов таблицы

mode – параметр работы функции: создание маски таблицы или ее столбцов

Функция записывает маски изображений в отдельные папки.

Примеры масок для изображения:





Теперь изображения, маски и размеры изображений сохраняются в отдельный датафрейм и записываются в csv файл для последующего удобного обращения к ним. Для этого используется функция compute_masks.

def compute_masks(images_dir):

images_dir – путь к папке с изображениями

Функция записывает и сохраняет csv файл с информацией об изображениях

## Описание модели и обучение.

Для обучения модели потребуются подготовленные ранее данные. Для этого считываем их из созданного csv файла в переменную train_df.

Создаем tensorflow-датасет для последующей загрузки в модель в переменной dataset.

Для подготовки изображений используется функция _parse_function.

def _parse_function(image, mask, colmask):

image – путь к изображению

mask – путь к маске с таблицей

colmask – путь к маске со столбцом

Загружает изображения в память, нормализует их, изменяет размер

Возвращает изображения и словарь с масками.

Создаем датасет для передачи модели для обучения, сохраняем его в переменную train_dataset.

Для проектирования и обучения модели машинного обучения используется библиотека TensorFlow.

Для увеличения эффективности модели используется заранее пред обученная модель компьютерного зрения “DenseNet”. Исключение (или dropout) нейронной сети равно 0.6 для избежания переобучения.

Определение слоев нейронной сети задается классами TableConvLayer(Layer) и ColumnConvLayer(Layer)

Создание модели происходит в функции build_tablenet.

```
def build_tablenet():
    Функция загружает выбранную пред обученную модель DenseNet и
    определяет структуру слоев.
    return model
```

Запуск обучения модели:

```
model.fit(
   train_dataset, epochs=100,
```

```
    steps_per_epoch= STEPS_PER_EPOCH,
    validation_data=val_dataset,
    validation_steps= VALIDATION_STEPS,
    callbacks=callbacks
)
```

## Применение модели к фотографиям для распознавания таблиц.

Теперь, когда нейронная сеть обучена для распознавания таблиц нужно применить ее на конкретной фотографии.

Для этого загружаем изображение с помощью функции из библиотеки PIL Image.open(image_filename), нормализуем его, изменяем размер.

С помощью модели получаем маску изображения, на которой показано расположение таблицы:

```
tab_mask, _ = model.predict(np_image)
```

Находим границы прямоугольника по этой маске и вырезаем его по этим координатам на исходном изображении. Таким образом мы вырезали из всей фотографии лишь нужную нам часть – таблицу.

```
image_orig = image_orig
x, y, w, h = cv2.boundingRect(tab_mask)
tab = image_orig.crop((x, y, x + w, y + h))
```

Теперь к выделенной части фотографии применяем функцию из библиотеки pytesseract – image_to_string(tab), которая считывает текст с изображения и возвращает строку с текстом.

```
text = pytesseract.image_to_string(tab)
```

Для последующей работы с текстом используется функция text_to_ecxel. Далее приведено описание ее работы.

1. В полученной строке избавляемся от лишних символом в начале и конце. Затем делим ее на несколько строк, соответствующих строкам таблицы с помощью метода split() используя в качестве разделителя “\n”.

2. Каждую полученную строку делим на слова, аналогичным образом, используя в качестве разделителя пробел.

3. Зная общую структуру протокола, мы можем записать полученные данные в нужные нам ячейки.

4. Полученное разбиение на ячейки организуем в датафрейм и записываем в эксель файл.

## Результаты.

### Эффективность различных подходов.

В работе представлено два алгоритма решения поставленной задачи:

1. Использование OpenCV для считывания структуры таблицы и PyTesseract для считывания данных

2. Использование нейронной сети, обученной с помощью TensorFlow и Keras и считывание данных с помощью PyTesseract

Применив эти алгоритмы к тестовым фотографиям, мы получили следующие результаты.

OpenCV + PyTesseract.

Эффективность работы с печатными данными:

0 - всевозможные конфигурации параметров функции image_to_string() не позволили распознать крестики и нолики (записанные буквами X и O соответственно)

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | X | O | | O | | | O | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | X | O | | O | | | X | O | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | O | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | O | | | O | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | O | | | X | X | X | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | X | O | | X | X | O | X | O | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | O | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | O | | | X | O | | X | X | X |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
| 1 | 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | | | | | | | | | |
| 2 | 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | | | | | | | | | |
| 3 | 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | | | |
| 4 | 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | | | | | | |
| 5 | 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | | | | | | | | | |
| 6 | 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | | | | | | | | | |
| 7 | 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | | | |
| 8 | 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | | | | | | | | | |

1 - эффективность 100% - вместо крестиков и ноликов было решено использовать плюсики и минусы, их система способна распознать

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | - | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | - | - | + | - | + | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
| 1 | 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | | + | | + | | | + | | |
| 2 | 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | | + | | + | | | - | + | |
| 3 | 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 4 | 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 5 | 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | | | | |
| 6 | 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | | + | | - | - | + | - | + | |
| 7 | 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 8 | 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | | - |

2 - эффективность 97.2% - 3 ячейки считаны неправильно

| Номер | Фамилия, имя | Дата рожд | | Организация | Нагр номер | 1 | 2 | 3 | 4 | Резул | Место |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | 4.15 | 4.01 | 4.08 | 3.80 | 4.15 | 4 |
| 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | 3.34 | 3.18 | 3.13 | 3.40 | 3.40 | 7 |
| 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | 3.20 | 2.95 | 3.25 | 3.25 | 3.25 | 8 |
| 4 | Гезина Екатерина | 02.12.2009 | 1ю | МГФСО | 761 | 3.74 | 4.14 | 4.24 | - | 4.24 | 3 |
| 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | 4.38 | 4.37 | 4.32 | 4.14 | 4.38 | 1 |
| 6 | Козлова Мария | 25.02.2010 | | | 29 | 3.58 | 3.59 | 3.41 | 3.60 | 3.60 | 5 |
| 7 | Кондратьева Софья | 12.01.2009 | 1ю | | 81 | 3.48 | 3.44 | 3.35 | 3.24 | 3.48 | 6 |
| 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | 3.95 | 4.30 | 4.29 | 4.10 | 4.30 | 2 |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Номер | Фамилия, имя | Дата рожд | | Организация | Нагр номер! | 1 | 2 | | | Резул | Место |
| 1 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | 4.15 | 4.01 | 4.08 | 3.80 | 4.15 | 4 |
| 2 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | 3.34 | 3.18 | 3.13 | 3.40 | 3.40 | 7 |
| 3 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | 3.20 | 2.95 | 3.25 | 3.25 | 3.25 | 8 |
| 4 | 4 | Гезина Екатерина | 02.12.2009 | 1ю | МГФСО | 761 | 3.74 | 4.14 | 4.24 | - | 4.24 | 3 |
| 5 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | 4.38 | 4.37 | 4.32 | 4.14 | 4.38 | 1 |
| 6 | 6 | Козлова Мария | 25.02.2010 | | | 29 | 3.58 | 3.59 | 3.41 | 3.60 | 3.60 | 5 |
| 7 | 7 | Кондратьева Софья | 12.01.2009 | 1ю | | 81 | 3.48 | 3.44 | 3.35 | 3.24 | 3.48 | 6 |
| 8 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | 3.95 | 4.30 | 4.29 | 4.10 | 4.30 | 2 |

Эффективность работы с данными, введенными с графического планшета:

1 - эффективность 98.6% - 2 ячейки не считаны (максимальная эффективность, сделано перебором рукописных символов)

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | - | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | - | - | + | - | + | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
| 1 | 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 2 | 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 3 | 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 4 | 189 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 5 | 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | | - | - | | | |
| 6 | 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | | - | + | - | + | |
| 7 | 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 8 | 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

2 - эффективность 62.9% - 2 ячейки не считаны (не улучшалось содержимое до хорошей эффективности)

| Номер | Фамилия, имя | Дата рожд | | Организация | Нагр номер | 1 | 2 | 3 | 4 | Резул | Место |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | 4.15 | 4.01 | 4.08 | 3.80 | 4.15 | 4 |
| 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | 3.34 | 3.18 | 3.13 | 3.40 | 3.40 | 7 |
| 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | 3.20 | 2.95 | 3.25 | 3.25 | 3.25 | 8 |
| 4 | Гезина Екатерина | 02.12.2009 | 1ю | МГФСО | 761 | 3.74 | 4.14 | 4.24 | - | 4.24 | 3 |
| 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | 4.38 | 4.37 | 4.32 | 4.14 | 4.38 | 1 |
| 6 | Козлова Мария | 25.02.2010 | | | 29 | 3.58 | 3.59 | 3.41 | 3.60 | 3.60 | 5 |
| 7 | Кондратьева Софья | 12.01.2009 | 1ю | | 81 | 3.48 | 3.44 | 3.35 | 3.24 | 3.48 | 6 |
| 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | 3.95 | 4.30 | 4.29 | 4.10 | 4.30 | 2 |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Номер | Фамилия, имя | Дата рожд | | Организация | Нагр номер} | 1 | 2 | 8 | 4 | Резул | Место |
| 1 | 1 | Асеева Серафима \ | 06.04.2010 | | МГФСО | 563 | “.15 | Ч.01 | 48 | . 80 | 415 | 5 |
| 2 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | 334 | 318 | 3.13 | 3.y0 | 3.40 | 7 |
| 3 | 3 | Галингер Арина \ | 24.03.2009 | 1ю | МГФСО | 125 | 3.20 | 295 | 3.25 | 325 | 3.25 | |
| 4 | 4 | Гезина Екатерина \ | 02.12.2009 | 1ю | МГФСО | 761 | 3. 74 | 19 | Ч.2.Ч | - | 4-24 | 3 |
| 5 | 5 | Зенченко Виктория \| | 17.01.2010 | | МГФСО | 83 | 433 | 4.37 | 4.32 | Ч.14Ч | 4-33 | 1 |
| 6 | 6 | Козлова Мария | 25.02.2010° | | | 29 | 358 | 3.56° | 394 | 3.60 ° | 36 | 5 |
| 7 | 7 | Кондратьева Софья \ | 12.01.2009 | 1ю | | 81 | 348 | 3-4 | 335 | 3.29 | 37 | 6 |
| 8 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | 3.95 | “.30 | 29 | “.40 | 4-3 | 2 |

Эффективность работы с данными, заполненными от руки:

57% эффективности

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | - | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | - | - | + | - | + | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | — 170 | \| 470 | 170 | 180 \| | — 180 | _ 180 | 0190 \| | | |
| 1 | 4179 | 3 | Асеева Серафима \| | 06.04.2010 | | МГФСО \| | 0563 \| | | + | | + | | | + | | |
| 2 | 179 | й \| | Ахмедова Амиль | '01.02.2011 | | МГФСО | 86 | | : | | + | | | - | + | |
| 3 | 19 | | Галингер Ариng } | 24.03.2009 | 1 | МГФСО _ | 425 | | | | | 1 | | + | | |
| 4 | 189 | 4 | Гезина Екатерина ] | 02.12.2009 ` | | МГФСО | 764 | | | | 9 | 3 | | + | | 3 |
| 5 | 4770 \| | 5 | ,Зенченко Виктория | `17.01.2010 | 5 | | 83 | + | | | - | | | - | | |
| 6 | 47 | 6 | _ Козлова Мария | 25.02.2010 | | ° МГФСО | | - | + | | - | - | + | | 8 | |
| 7 | 139 | 7 | Кондратьева Софья} | 12.01.2009` | 8 | МГФСО | 84 | | | | | 1 | | 9 | | |
| 8 | кч` | 8 | Короткова Ульяна | 05.06.2010. | 7 | МГФСО | | + | | | - | 38 | - | | | |

Вывод: Программа, написанная с полным опором на модуль OpenCV показала внушительный результат. Она идеально считала структуру, приведенных таблиц и с хорошей эффективностью смогла считать содержимое таблицы. Печатные символы считываются идеально. А рукописные символы, записанные на графическом планшете, считываются немного лучше, чем те, что пишутся ручкой на распечатке с дальнейшим сканированием. Также примечательно то, что рукописи можно довести до такого состояния, чтобы программа могла их считывать почти со стопроцентной эффективностью.

TensorFlow + PyTesseract.

Эффективность работы с печатными данными:

0

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | X | O | | O | | | O | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | X | O | | O | | | X | O | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | O | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | O | | | O | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | O | | | X | X | X | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | X | O | | X | X | O | X | O | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | O | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | O | | | X | O | | X | X | X |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ач | выс\|Номер\| | _ | Фамилия, | имя | _ | \| | Дата | рожд |
| 1 | 170 | - | - | - | - | - | - | - | - |
| 2 | 170 | - | - | - | - | - | - | - | - |
| 3 | 190 | - | - | - | - | - | - | - | - |
| 4 | 180 | - | - | - | - | - | - | - | - |
| 5 | 170 | - | - | - | - | - | - | - | - |
| 6 | 170 | - | - | - | - | - | - | - | - |
| 7 | 190 | - | - | - | - | - | - | - | - |
| 8 | 170 | - | - | - | - | - | - | - | - |
| 9 | [ | + | \| | Асеева | Серафима | 06042010 | \| | - | - |
| 10 | [ | 2 | \| | Ахмедова | Амиль_01.02.2011\|_—_\| | - | - | - | - |
| 11 | [а | \| | гезина | Екатерина | \|02.12.2009]_—_\| | - | - | - | - |
| 12 | [ | 5 | \|Зенченко | Виктория | 17.01.2010]\|_—_\| | - | - | - | - |
| 13 | [ | 6 | \| | Козлова | Мария_\|25.02.2010\|___\| | - | - | - | - |
| 14 | [ | 7 | \|кондратьева | Софья12.01.2009\|_3 | _ | - | - | - | - |
| 15 | [ | в | ] | Короткова | Ульяна | [05.06.2010] | — | ] | - |
| 16 | МГ | - | - | - | - | - | - | - | - |
| 17 | МГ | - | - | - | - | - | - | - | - |
| 18 | МГ | - | - | - | - | - | - | - | - |
| 19 | МГ | - | - | - | - | - | - | - | - |
| 20 | МГ | - | - | - | - | - | - | - | - |
| 21 | МГ | - | - | - | - | - | - | - | - |
| 22 | МГ | - | - | - | - | - | - | - | - |
| 23 | МГ | - | - | - | - | - | - | - | - |

1

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | - | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | - | - | + | - | + | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | [а | [ | Талингер | Арина_\|24.03.2009] | ю | [ | мгосо | 5 | [ | [ | - | - | - | - | - | - | - | - | - |
| 1 | [а | \| | гезина | Екатерина | \|02.12.2009] | [ | мгесо | [ | 761 | [ | — | \| | Opp | - | - | - | - | - | - |
| 2 | [ | 5 | \|Зенченко | Виктория17.01.2010\|_ | [ | мгесо | [ | ва | ] | + | [ | ] | [ | - | \| | - | \| | - | р |
| 3 | 6[ | козлова | мавия_[25ог-озо[ | — | [ | мгесо | [ | [ | [ | [ | [ | [ | [ | - | - | - | - | - | - |
| 4 | -- _-------- | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 5 | Короткова | Ульяна | \| | 05.06.2010 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

2

| Номер | Фамилия, имя | Дата рожд | | Организация | Нагр номер | 1 | 2 | 3 | 4 | Резул | Место |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | 4.15 | 4.01 | 4.08 | 3.80 | 4.15 | 4 |
| 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | 3.34 | 3.18 | 3.13 | 3.40 | 3.40 | 7 |
| 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | 3.20 | 2.95 | 3.25 | 3.25 | 3.25 | 8 |
| 4 | Гезина Екатерина | 02.12.2009 | 1ю | МГФСО | 761 | 3.74 | 4.14 | 4.24 | - | 4.24 | 3 |
| 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | 4.38 | 4.37 | 4.32 | 4.14 | 4.38 | 1 |
| 6 | Козлова Мария | 25.02.2010 | | | 29 | 3.58 | 3.59 | 3.41 | 3.60 | 3.60 | 5 |
| 7 | Кондратьева Софья | 12.01.2009 | 1ю | | 81 | 3.48 | 3.44 | 3.35 | 3.24 | 3.48 | 6 |
| 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | 3.95 | 4.30 | 4.29 | 4.10 | 4.30 | 2 |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | [ | 3 | Талингер | Арина | _ | [24.03.2009\|Г | 19 | — | [Мг |
| 1 | [ | _а | \| | гезина | Екатерина | 02.12.2009\| | мг | - | - |
| 2 | [ | 5 | \|Зенченко | Виктория17.01.2010] | _ | [ | мк | - | - |
| 3 | [ | 6 | \| | Козлова | Мария_\|25.02.2010]___ | [ | мк | - | - |
| 4 | [ | 7 | \|кондратьева | Софья12.01.2009] | 3 | _ | [ | мк | - |
| 5 | 8 | Короткова | Ульяна | \| | 05.06.2010. | МГ | - | - | - |

Эффективность работы с данными, введенными с графического планшета:

1

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | - | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | - | - | + | - | + | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | [ | 3 | Талингер | Арина | _ | [24.03.2009\|Г | 19 | — | [Мг |
| 1 | [ | _а | \| | гезина | Екатерина | 02.12.2009\| | мг | - | - |
| 2 | [ | 5 | \|Зенченко | Виктория17.01.2010] | _ | [ | мк | - | - |
| 3 | [ | 6 | \| | Козлова | Мария_\|25.02.2010]___ | [ | мк | - | - |
| 4 | [ | 7 | \|кондратьева | Софья12.01.2009] | 3 | _ | [ | мк | - |
| 5 | 8 | Короткова | Ульяна | \| | 05.06.2010. | МГ | - | - | - |

2

| Номер | Фамилия, имя | Дата рожд | | Организация | Нагр номер | 1 | 2 | 3 | 4 | Резул | Место |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | 4.15 | 4.01 | 4.08 | 3.80 | 4.15 | 4 |
| 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | 3.34 | 3.18 | 3.13 | 3.40 | 3.40 | 7 |
| 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | 3.20 | 2.95 | 3.25 | 3.25 | 3.25 | 8 |
| 4 | Гезина Екатерина | 02.12.2009 | 1ю | МГФСО | 761 | 3.74 | 4.14 | 4.24 | – | 4.24 | 3 |
| 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | 4.38 | 4.37 | 4.32 | 4.14 | 4.38 | 1 |
| 6 | Козлова Мария | 25.02.2010 | | | 29 | 3.58 | 3.59 | 3.41 | 3.60 | 3.60 | 5 |
| 7 | Кондратьева Софья | 12.01.2009 | 1ю | | 81 | 3.48 | 3.44 | 3.35 | 3.24 | 3.48 | 6 |
| 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | 3.95 | 4.30 | 4.29 | 4.10 | 4.30 | 2 |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 0 | Фамилия, | имя | Дата | рожд, | Организаци | - | - | - |
| 1 | [ | Асеева | Серафима_\|06.04.2010 | _ | [ | мгФСО | - | - |
| 2 | [ | Ахмедова | Амиль_\|01.02.2011 | [ | [ | мгФСО | - | - |
| 3 | МгФСО | - | - | - | - | - | - | - |
| 4 | МгФСО | - | - | - | - | - | - | - |
| 5 | [ | Зенченко | Виктория | [ | 17.01.2010 | _ | [ | мгФСО |
| 6 | [ | _Козлова | Мария | _[ | 25.02.2010 | [ | \| | - |
| 7 | Короткова | Ульяна | \| | 05.06.2010 | МГФСО | - | - | - |

Эффективность работы с данными, заполненными от руки:

| Нач выс | Номер | Фамилия, имя | Дата рожд | | | Нагр номер | 170 | 170 | 170 | 180 | 180 | 180 | 190 | 190 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 170 | 1 | Асеева Серафима | 06.04.2010 | | МГФСО | 563 | - | + | | + | | | + | | |
| 170 | 2 | Ахмедова Амиль | 01.02.2011 | | МГФСО | 86 | - | + | | + | | | - | + | |
| 190 | 3 | Галингер Арина | 24.03.2009 | 1ю | МГФСО | 125 | | | | | | | + | | |
| 180 | 4 | Гезина Екатерина | 02.12.2009 | | МГФСО | 761 | | | | + | | | + | | |
| 170 | 5 | Зенченко Виктория | 17.01.2010 | | МГФСО | 83 | + | | | - | - | - | | | |
| 170 | 6 | Козлова Мария | 25.02.2010 | | МГФСО | 29 | - | + | | - | - | + | - | + | |
| 190 | 7 | Кондратьева Софья | 12.01.2009 | 3 | МГФСО | 81 | | | | | | | + | | |
| 170 | 8 | Короткова Ульяна | 05.06.2010 | | МГФСО | 57 | + | | | - | + | | - | - | - |

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | ›роткова | ульяна | \| | Уэ,чо.еча |

Вывод:

При обработке изображений с помощью алгоритма на Keras/TensorFlow теряется структура изображения и нет смысла говорить об эффективности. Данные, заполненные от руки не считались ни в одном случае.

Общий вывод и выбор наилучшего подхода(наилучшей методологии):

## Заключение.

В данном проекте были рассмотрены задачи по разработке программы по распознаванию табличных протоколов различных соревнований для их перевода из с бумажных носителей в excel таблицы. В процессе разработки главную роль играли такие библиотеки, как OpenCV, PyTesseract, TensorFlow и Keras.

Были получены следующие результаты:

- Разработан алгоритм распознавания с использованием OpenCV + PyTesseract;
- Разработан алгоритм распознавания с использованием Keras/TensorFlow + PyTesseract;
- Проведено сопоставление эффективности работы двух алгоритмов на основе разных инструментов;
- Определены наилучшие методологии работы распознавания на основе используемых инструментов.

**Будущее проекта.**

Проект имеет внушительные перспективы по доработке и улучшению качества точности распознавания. В будущем возможна доработка алгоритма детекции разметки таблицы при помощи технологий OpenCV, а также проектирование собственной нейронной сети, обученной на специализированном датасете в качестве замены PyTesseract. А также при глубинном изучении возможностей TensorFlow возможно значительное улучшение работы программы при координальной её доработке и составлении новых тренировочных данных для модели распознавания таблиц.