НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Майнор Бизнес-Информатика
Современные информационные технологии в бизнесе

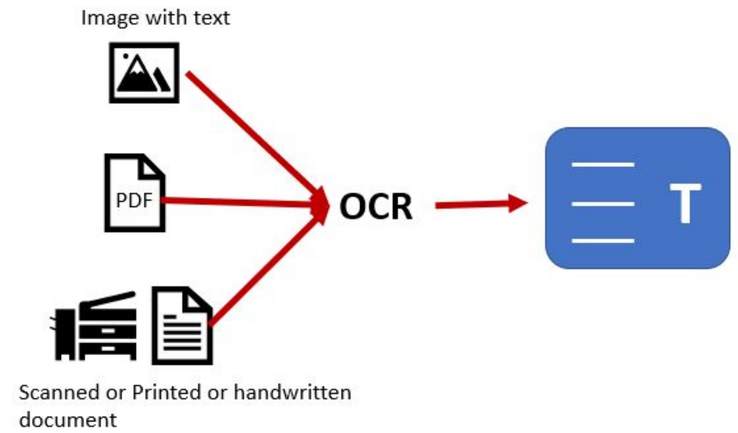
Оптическое распознавание символов (OCR). Алгоритмы удаление мусора

Выполнил студент 4-й группы
Индюченко Никита

Москва, 2022

Что такое OCR

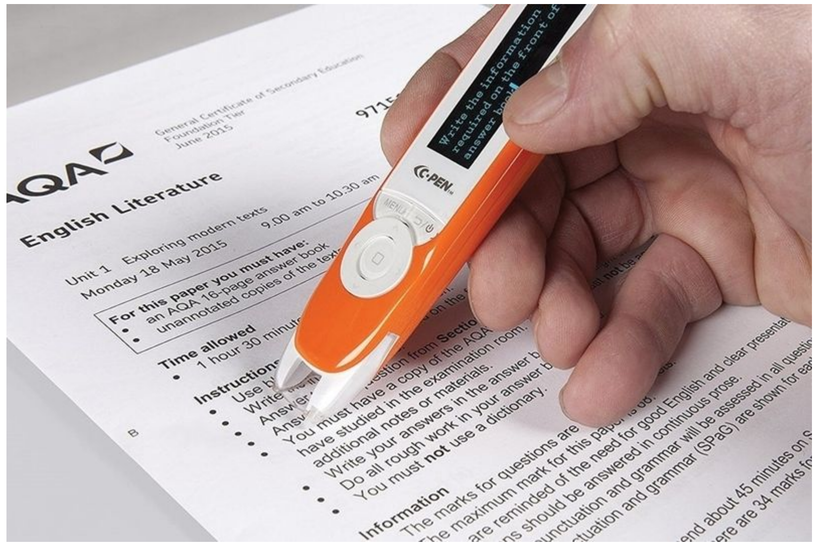
Оптическое распознавание символов - это механическое или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные, использующие для представления символов в компьютере



История оптического распознавание символов

В 1978 году компания “Курцвейл компьютер продактс” начала продажи первой коммерчески успешной компьютерной программы оптического распознавания символов. Два года спустя Курцвейл продал свою компанию корпорации “Ксерокс”.

Первой программой, распознающей кириллицу, была программа **«AutoR»** российской компании «ОКРУС». Программа начала распространяться в 1992 году. 1993 году вышла технология распознавания текстов российской компании **ABBYY**



PenReader - устройство, сканирующая печатный текст, созданный компанией ABBYY

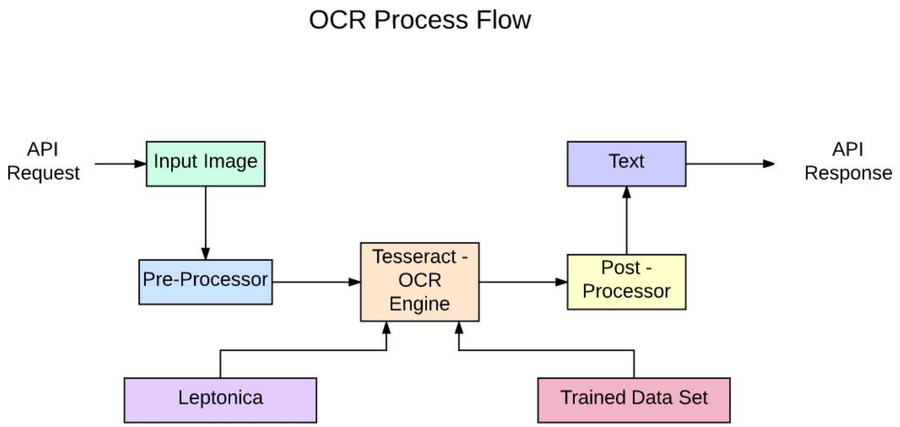
Источники: Wikipedia.org, Nirmech.ru,

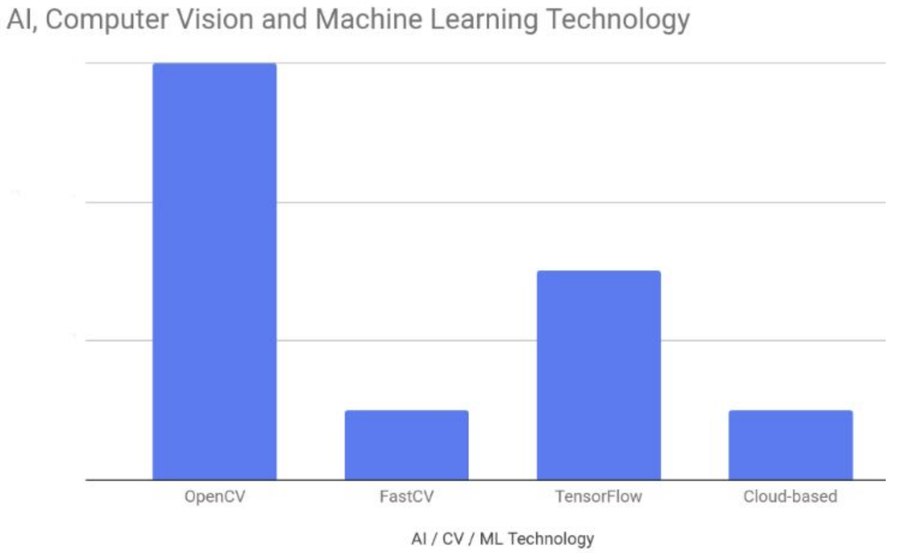
Распознавание протоколов

В 2022 году студентам 1-ого курса ОП “Прикладная математика” была предложена идея по созданию программы, которая должна конвертировать графические файлы в excel таблицы

Основные инструменты для работы с изображением

- PIL
- OpenCV
- PyTesseract
- Tensorflow
- CUDA
- TPU и Colaboratory
- Keras





Источники: dzone.com, nanonets,com

Задачи и первая реализация нейронной сети

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

2 - **Основной алгоритм** (OpenCV+PyTesseract).

- Разбиение таблицы на ячейки.
- Распознавание содержимого каждой ячейки

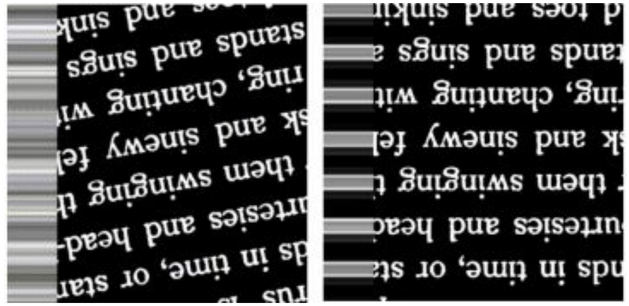
Шаги разбиения таблицы

1 - Распознание вертикальных и горизонтальных линий

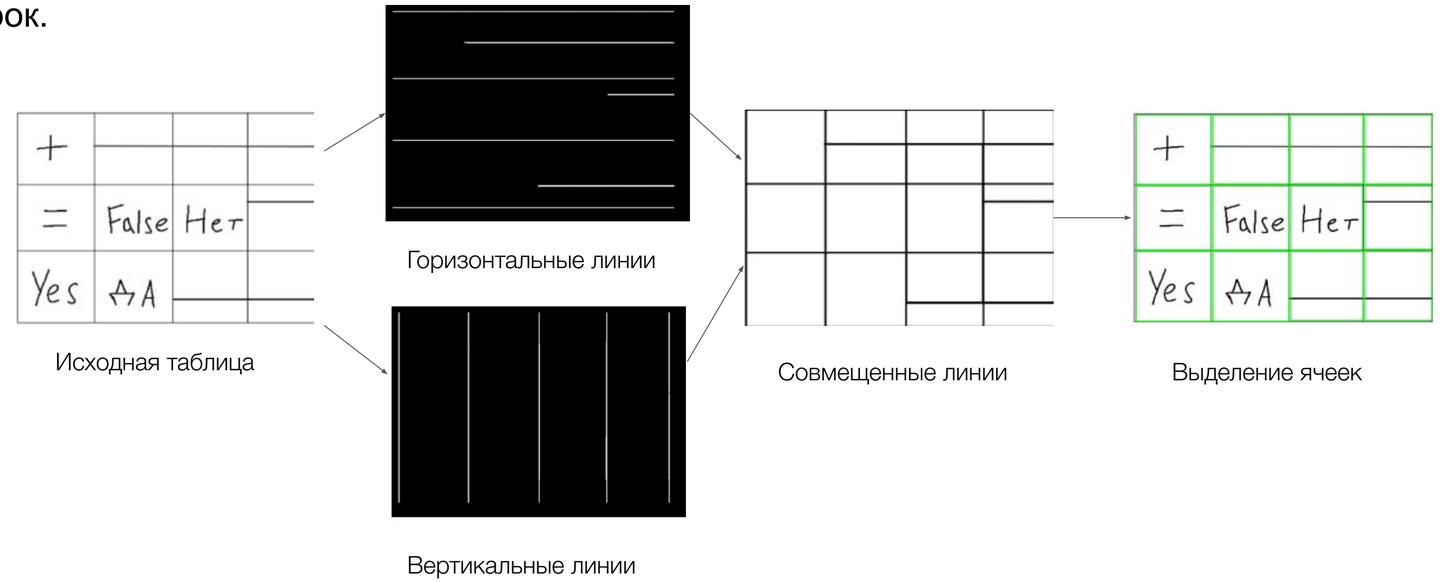
2 - Совмещение

3 - Определение “ложных” линий

4 - Выделение ячеек



Фотография с наклоном и без наклона



Исходная таблица

Горизонтальные линии

Вертикальные линии

Совмещенные линии

Выделение ячеек

Источники: zephyrnet.com

Распознавание символов

Сравнение работоспособности TensorFlow и OpenCV

OpenCV + PyTesseract

Эффективность работы с печатными данными:

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	+			-	+		-	-	-

Переделанный исходник: крестики и нолики заменили на + и - 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

Исходник

	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									

Результат

	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	+		-	+		-		-	

Результат с “новым” исходником

Выводы:

- PyTesseract не очень богат алфавитом различных символов, так как он не распознаёт такие знаки, как крестики и нолики
- Заменив эти символы на + и -, проблема со сканированием данных решается
- Печатные данные распознаёт со 100% эффективностью

Источники: отчёт по проектной работе “Распознавание протоколов”, zephyrnet.com

Продолжение

Эффективность работы с данными, введенными с графического планшета:

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	+			-	+		-	-	-

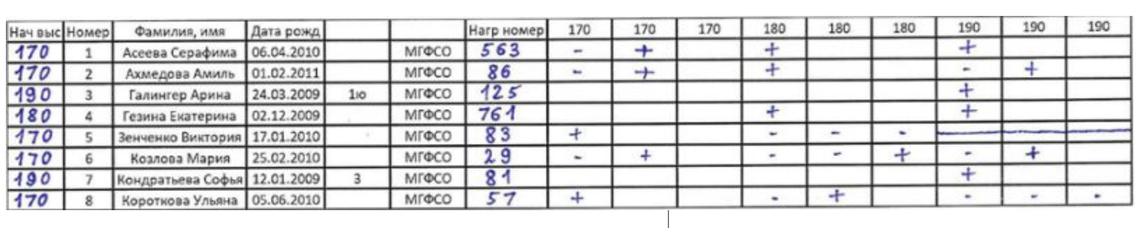
Исходник

	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	+			-	+		-	-	-

Результат

Эффективность 98,6% - 2 ячейки не считаны

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	+			-	+		-	-	-

Исходник

	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 Аринг }	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	-			

Результат

Эффективность 57%

Источники: отчёт по проектной работе “Распознавание протоколов”

TensorFlow+PyTesseract

Эффективность работы с печатными данными:

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

Исходник

	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	-	-	-	-	-	-	-	-

Результат

Эффективность работы с данными, введенными с графического планшета:

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	+			-	+		-	-	-

Исходник

	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	-	-	-

Результат

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.

Источники: отчёт по проектной работе “Распознавание протоколов”

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Спасибо за внимание