НЕТОЛОГИЯ

Data Scientist: расширенный курс

Пояснительная записка к проекту

на тему:

«Детекция объектов на видео и отслеживание проникновения в запретную зону»

Автор - Дементьева Мария

Группа DSU-74

2026

Содержание

1. Введение и постановка задачи

Данная работа выполняется в интересах владельца частного дома, желающего обеспечить безопасность своего жилья и прилегающей к нему территории без необходимости постоянного вмешательства и контроля ложных тревог. Это особенно важно в случаях, когда владелец находится в отъезде и не имеет возможности оперативно приехать для проверки ситуации.

В настоящее время на придомовой территории установлено видеонаблюдение. Однако сигнал о проникновении на участок часто срабатывает при любом движении в зоне камеры — будь то движение животных, птиц или даже падающий снег, который находится близко к объективу. Это приводит к частым ложным срабатываниям и создает ненужные тревоги.

Основная цель проекта— автоматизировать процесс определения появления посторонних лиц, исключая при этом ошибочные срабатывания из-за животных, птиц или других несущественных объектов.

Кроме того, важно, чтобы система могла отличать владельцев от посторонних лиц. В случае, если на территорию заходит именно владелец — система не должна подавать тревогу, а при обнаружении постороннего — необходимо своевременно зафиксировать и оповестить. При этом не нужно отслеживать перемещения объекта или нарушение границ участка, а просто фиксировать факт проникновения, т.к. камера установлена таким образом, что не захватывает территорию за пределами участка.

Для этой цели использована модель нейронной сети, обученная распознавать классы объектов: "владелец" и "люди" (посторонние).

Метриками оценки качества модели являются Precision (точность), которая показывает долю правильно распознанных случаев среди всех срабатываний системы, и Recall (полнота), отражающая, сколько истинных объектов класса система смогла обнаружить. Вместе с ними используются метрики mAP50 и mAP50-95 (Mean Average Precision) — они измеряют качество обнаружения объектов при разных порогах метрики IoU ((Intersection over Union - насколько точно предсказаны рамки объекта) и дают целостную картину эффективности модели.

Учитывая тот факт, что нужно отслеживать в основном факт правильного обнаружения и классификации объектов (посторонние или владельцы), ключевыми метриками являются Recall (полнота) и mAP50-95. Высокий показатель Recall свидетельствует о том, что система практически не пропускает объекты, что особенно важно в случаях, когда пропуск даже одного объекта может иметь серьезные последствия. mAP50-95 используется для достижения баланса между обнаружением всех объектов и минимизацией ложных тревог, т.к. она объединяет показатели точности, полноты и локализации объектов.

2. Описание данных и их особенности

За базовый набор данных выбран датасет **thief detection dataset**
https://www.kaggle.com/datasets/janstylewis7/improvedthiefdetectiondataset

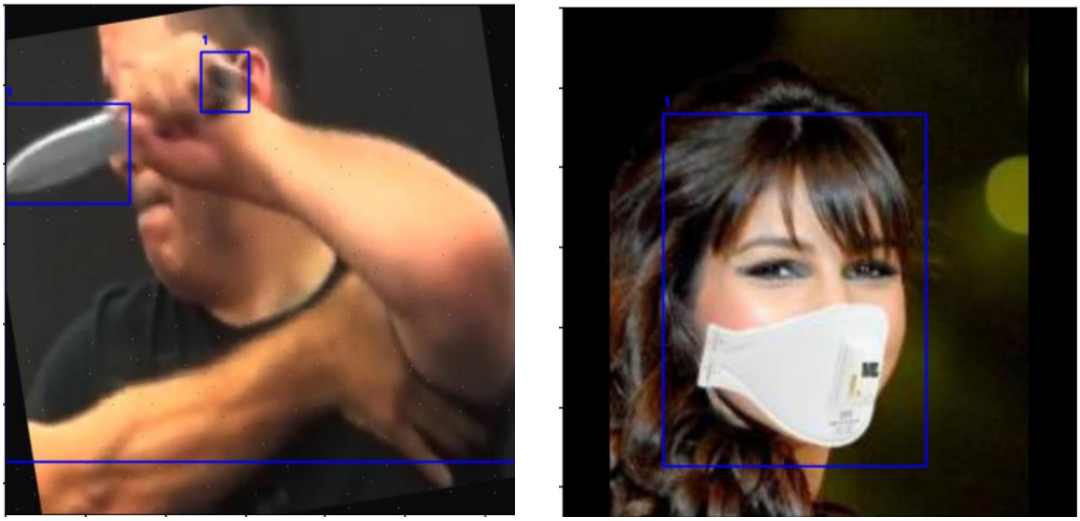
Набор данных предназначен для обучения моделей обнаружения объектов с целью идентификации людей и/или их подозрительного поведения на видеозаписях с камер наблюдения. Данные состоят из аннотированных кадров, извлеченных из различных сценариев, имитирующих реальные кражи и подозрительную активность. Аннотации выполнены в формате моделей YOLO.

В исходном датасете всего 12 480 кадров с разметкой на 2 класса:
0 - человек - просто люди;
1 - подозреваемые - люди в масках, ножи, люди с подозрительным поведением, люди с оружием.

Пустых кадров без объектов и разметки нет. Все изображения в наборе разного размера.
На рис.1 представлены примеры кадров исходного датасета.

Рис.1 Примеры кадров исходного датасета
а) класс 0 “человек”

б) класс 1 “подозреваемые”



Количество и распределение объектов классов приведено в Таблице 1.

Табл.1. Распределение объектов разных классов

Тип кадра	**Количество кадров**	**Количество людей**	**Количество подозреваемых**
Кадры только с людьми (класс 0)	5339	13029	0
Кадры только с подозреваемыми (класс 1)	4 613	0	5 438
Смешанные кадры (классы 0 и 1)	2 528	6 093	6 299
Итого:	12 480	19 122	11 737

Судя по полученным цифрам количество объектов каждого класса сбалансировано и нет совсем большого перекоса в сторону какого-то одного класса.

Т.к. по условию задачи необходимо распознать попадание посторонних лиц на территорию частного дома, то датасет был дополнен кадрами владельцев чтобы обученная модель распознавала не только посторонних, но и владельцев.
Добавленным кадрам с владельцами в файлах разметки был присвоен класс 2 - Владельцы.

Рис.2 Примеры дополненных кадров с классом 2 (Владельцы)

Также датасет был дополнен пустыми кадрами без людей, подозреваемых и владельцев. Для таких кадров файлы разметки не создавались.

Было добавлено суммарно 516 кадров, из которых:

- 382 кадра с владельцами;
- 134 пустых кадра.

Добавленное количество кадров на 2-й класс (Владельцы) составляет всего порядка 3% от суммарного количества изображений на все 3 класса.

Если принять во внимание только кадры с людьми или кадры с подозреваемыми, то добавленное количество класса 2 (Владельцы) составляет порядка 7-8% от классов 0 или 1.

Учитывая то, что основной интерес вызывала работа моделей YOLO от Ultralytics, то были приняты во внимание рекомендации Ultralytics о стартовом количестве аннотированных объектов на 1 класс порядка нескольких сотен (ссылка на статью Ultralytics).

Исследование датасета выполнено в файле diplom_dataset.ipynb

3. Описание обработки данных и разделение на train, test, validation

Исходный датасет был разбит на:
train - 10 420 кадров (83% от общего количества)
test - 735 кадров (6% от общего количества)
valid - 1 325 кадров (11% от общего количества).

Для добавления кадров с объектами класса 2 (Владельцы) было снято нескольких видео в разной обстановке и в разной одежде. Затем из каждого видео автоматически был выбран каждый 10-й кадр, на который потом были нанесены рамки разметки для детекции объектов. Число 10 для сохранения каров было определено эмпирически чтобы кадры были не сильно похожи друг на друга и чтобы получить достаточное количество кадров.

Заполнение txt файлов разметки с номером класса и координатами рамок разметки для детекции выполнялись автоматически посредством сканирования цвета пикселей рамок и вычислением координат. Аннотации выполнялись аналогично исходному датасету, в формате моделей YOLO.

Параллельно проводилось распределение изображений по папкам train, test, valid. Все изображения, попадающие в папку train, были продублированы и повернуты на произвольный угол для увеличения количества объектов для обучения.

Распределение добавленных кадров:
train - 386 кадров (75% от общего количества добавленных кадров)
test - 65 кадров (12,5% от общего количества)
valid - 65 кадров (12,5% от общего количества).

Деление кадров на train, test, validation выполнялось случайным образом при помощи random.shuffle, затем отбиралось нужное количество в каждую папку.

По окончании формирования добавочного датасета с разметкой была проведена визуальная оценка качества сгенерированных файлов разметки для каждого набора.

Всего в итоговом датасете получилось:
train - 10 806 кадров
test - 800 кадров
valid - 1390 кадров

Обработка видео с нарезкой кадров, разделение на train, test, validation и формирование txt файлов разметки выполнено в файле diplom_add_dataset.ipynb

4. Описание возможных решений и планируемая архитектура

Задача автоматического распознавания объектов на видео является задачей детекции, области компьютерного зрения, которая включает в себя локализацию и выделение объекта рамками и определение его класса. Поэтому для ее решения будем применять часто используемые алгоритмы задач детекции на базе сверточных нейронных сетей.

Необходимо учитывать технические возможности - обучение есть возможность проводить только в среде google colab в бесплатном режиме.

Т.к. для обучения нейронной сети с нуля требуется большое количество размеченных изображений, то необходимо было выбрать уже предобученную на большом количестве данных модель и дообучить ее на подготовленном ранее датасете.

Выбор конкретной предобученной модели проводился среди наиболее часто применяемых для задач детекции в 2025 году моделей по информации из интернета. Были рассмотрены Yolo11 и RT-DETR.

Модель RT-DETR (Real-Time Detection Transformer) обнаружения объектов создана для использования в режиме реального времени.

RT-DETR использует гибридную архитектуру кодировщика-декодера. В ней используется CNN-база для извлечения признаков, которые затем обрабатываются кодировщиком-декодером на основе трансформера. Изначально представлялось, что это неплохой вариант для детекции объектов на видео, но при обучении модели выяснилось, что бесплатных ресурсов google colab недостаточно. Даже при уменьшении размеров изображения до 128x128 за 5.5 часов было выполнено всего 17% 1-й эпохи. Характеристики RT-DETR приведены в таблице 2.

Таблица 2. Характеристики модели RT-DETR

Модель	**mAP val 50-95**	**Кол-во параметров (млн)**	**Сложность GFLOPS**
RT-DETR-L	53,4	76	259

Следующей рассмотрена линейка моделей YOLO11 (You Only Look Once) на базе сверточных слоев (CNN). Главное преимущество архитектуры YOLO11 в данной ситуации это ее вычислительная эффективность в сочетании с высокой точностью.

В линейку YOLO11 входит 5 вариантов нейросетей, различающихся числом параметров и, соответственно, точностью и скоростью исполнения/обучения.

Характеристики моделей YOLO11 приведены в таблице 3.

Таблица 3. Характеристики моделей YOLO11

Модель	**mAP val 50-95**	**Кол-во параметров (млн)**	**Сложность GFLOPS**
YOLO11n	39.5	2.6	6.5
YOLO11s	47.0	9.4	21.5
YOLO11m	51.5	20.1	68.0
YOLO11l	53.4	25.3	86.9

Учитывая имеющиеся ограничения вычислительных ресурсов были выбраны модели Yolo.

Предполагается дообучить модель на подготовленном датасете для возможности распознавания не только людей, но и владельцев. Затем обрабатывать каждый 10-й кадр видео с помощью полученной модели для обнаружения и классификации объектов. Обработка не каждого кадра, а каждого 10-го предлагается с целью сокращения времени на работу модели. Объект на 10 кадров не скроется из зоны обзора камеры, следовательно нет смысла передавать в модель каждый кадр.

5. Описание обучения

По данным из интернета модель YOLO11 отлично справляется с детекцией объектов без дополнительного обучения. Было интересно сравнить точность обнаружения именно людей модели без дообучения и моделей обученных на подготовленном датасете.

Далее, с учетом ограничений вычислительных ресурсов, было проведено дообучение модели YOLO11n (Nano) с самым маленьким количеством параметров. Размеры изображений в датасете варьировались в пределах от 416 до 1800 пикселей, квадратной или прямоугольной формы.

Для возможности обучения YOLO11n пришлось уменьшить размер изображения до 128x128 пикселей.

Обучение проводилось на 10 и 50 эпохах.

Также получилось выполнить обучение YOLO11s (small), но с с увеличением размера батча в 2 раза (32 вместо 16), с количеством эпох равным 10.

Обучение проводилось на train наборе данных с валидацией на valid наборе.

Test набор не использовался в процессе обучения. На нем проводилось итоговое сравнение моделей.

Остальные параметры приняты установленными в моделях по умолчанию и одинаковыми для всех моделей.

Результаты обучения моделей сведены в таблицы 4 и 5.

Таблица 4. Результаты обучения моделей на **Train**/Valid наборе в ходе обучения

Модель	**Время обучения, ч**	**Recall**		**MaP50-95**	
		Класс Люди	**Класс Владельцы**	**Класс Люди**	**Класс Владельцы**
yolo11n без дообучения	-	**0.428** 0,495	-	**0.121** 0,201	-
yolo11n дообученная 10 эпох	2	**0.572** 0,534	**0.884** 0,975	**0.375** 0,363	**0.54** 0,68
yolo11n дообученная 50 эпох	8,5	**0.62** 0,592	**0.918** 0,978	**0.436** 0,403	**0.56** 0,689
yolo11s дообученная 10 эпох	28	**0.713** 0,659	**0.925** 1	**0.495** 0,466	**0.695** 0,809

Таблица 5. Результаты обучения моделей на **Test** наборе

Модель	**Recall**		**MaP50-95**	
	Класс Люди	**Класс Владельцы**	**Класс Люди**	**Класс Владельцы**
yolo11n без дообучения	0,52	-	0,251	-
yolo11n дообученная 10 эпох	0,621	0,952	0,437	0,686
yolo11n дообученная 50 эпох	0,638	1	0,47	0,613
yolo11s дообученная 10 эпох	0.728	0.976	0.533	0.787

Из полученных результатов видно, что лучшие результаты показывает на **test** наборе показывает **yolo11s дообученная на 10** эпох

При этом, все дообученные модели классифицируют и локализуют объекты класса Люди лучше чем исходная модель.

Обучение моделей yolo11n проводилось в файле diplom_model.ipynb

Обучение модели yolo11s проводилась в файле model_test.ipynb .

Валидация всех моделей на test и train наборах проводилась в файле diplom_models_validation.ipynb

6. Описание итогового результата и, если есть, применение на практике

На основе сводных результатов обучения, в качестве финальной выберем **yolo11s 10 эпох.**

В результате обучения модели **yolo11s 10 эпох** были получены следующие ключевые показатели на **test** данных:

По классу Владельцы:

Recall : 0,976, что означает, что модель успешно обнаружила 97% объектов данного класса.

mAP50-95: 0,787 — что свидетельствует о точности обнаружения и классификации объектов выше средней, учитывая как расположение, так и правильность меток.

По классу Люди:

Recall : 0,728, что означает, что модель успешно обнаружила 72% объектов данного класса.

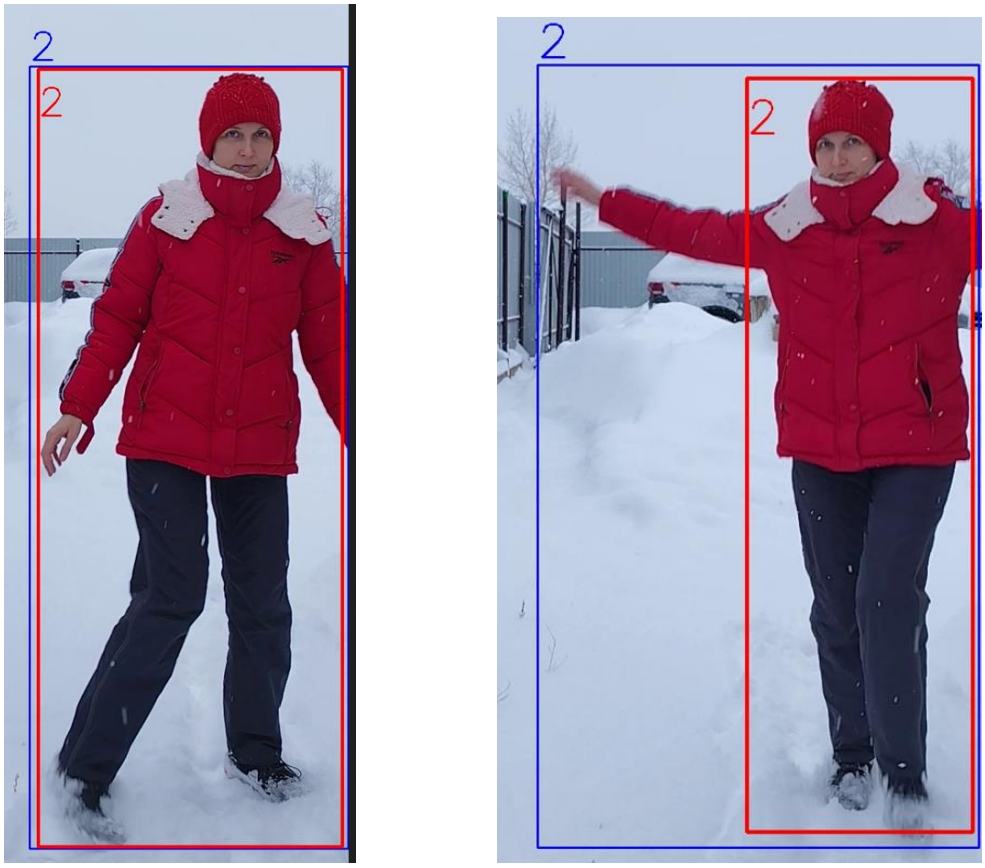
mAP50-95: 0,533 — что свидетельствует о средней точности обнаружения и классификации объектов, учитывая как расположение, так и правильность меток.

FLOPs: 21,3 GFLOPs — расчетные затраты на обучение модели.

Время обучения на cpu CPU (Intel Xeon CPU @ 2.20GHz) 28 часов.

На рис.3 представлены некоторые результаты детекции на кадрах из **test** набора. Визуализация детекции и истинной разметки на выборочных кадрах из test набора выполнена в файле diplom_dataset.ipynb .

Рис. 3. Выборочные результаты детекции на кадрах из test набора.
Синим - истинная аннотация, красным - предсказанная
Класс 0 - Люди, класс 2 - Владельцы



Выводы: Выбранная модель демонстрирует в среднем хороший уровень обнаружения объектов (recall=0,976 и 0,728), однако качество локализации и классификации объектов требует дальнейшего улучшения, что отражается в средней точности mAP (0.787 и 0.533).

Таким образом, модель обнаружит проникновение на территорию человека, но не всегда сможет классифицировать Владелец это или посторонний, что уменьшает, но не исключает полностью ложные срабатывания. При этом, исключены ложные срабатывания на несущественные объекты (животные, снег и т.д.).

Была проведена проверка работы выбранной модели на 2-х тестовых видео, которые не использовались при формировании датасета. 1-е видео с присутствием класса Владелец и собакой при нормальном освещении, 2-е с присутствием класса Владелец, класса Люди и собакой при плохом освещении.

Проверка проводилась в файле diplom_video.ipynb . Из-за конфликта google colab с cv2.imshow проверка тестового видео проводилась в среде VSCode на компьютере.

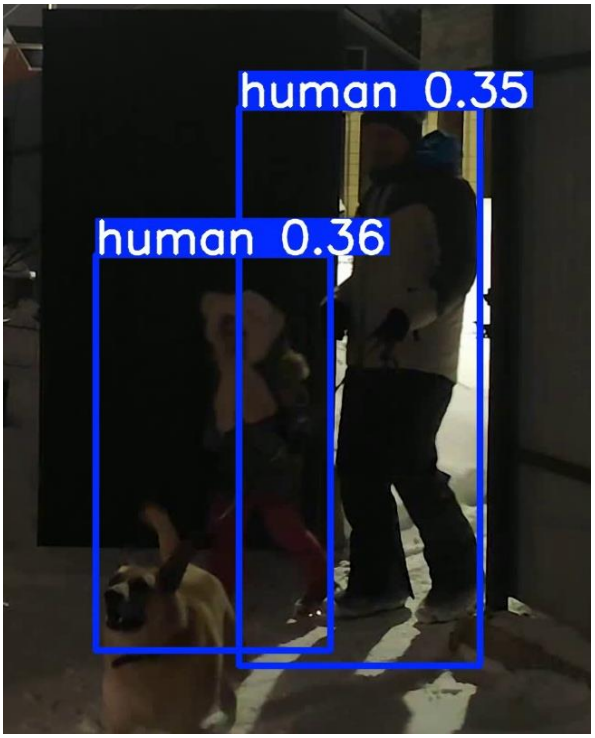
1-е тестовое видео с результатами детекции доступно по ссылке , 2-е по ссылке.

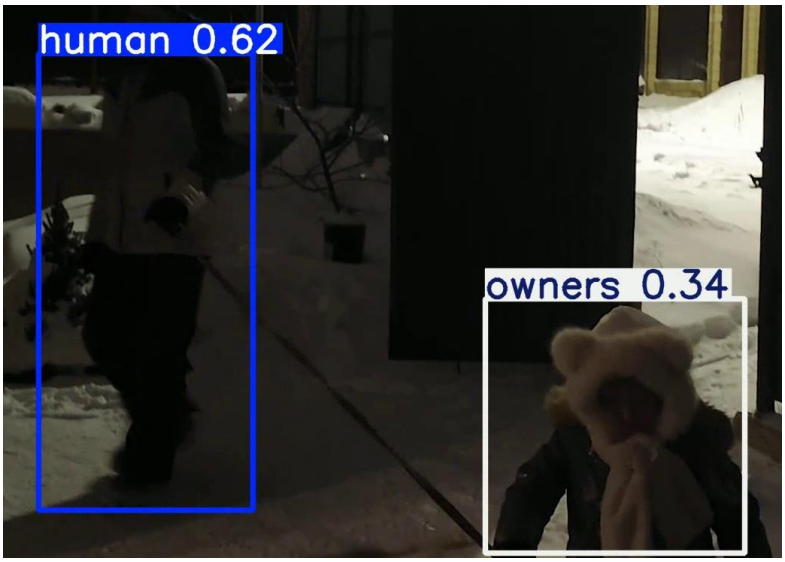
На рис. 4 представлены скриншоты из тестового видео с результатами детекции.

Рис.4. Кадры с результатами детекции из тестового видео

а) Видео 1 с присутствием класса Владелец и собакой при нормальном освещении

б) видео 2 с присутствием класса Владелец, класса Люди и собакой при плохом освещении





Из кадров видно, что модель не всегда находит на видео людей и владельцев, однако она также не реагирует на собаку.

Также, модель иногда один объект распознает дважды, а класс Владелец часто принимает за класс Люди.

Данные результаты получены при пороге уверенности conf = 0,25. Увеличение порога повлечет за собой увеличение кадров, в которых модель не обнаружит никаких объектов, а снижение приведет к ложным срабатываниям - собаку примет за человека.

7. Заключение с выводами и планами на дальнейшее развитие

На основании полученных результатов были сделаны следующие выводы:

1. Количества дополненных кадров с классом 2 (Владельцы) слишком мало для качественного обучения модели. Нужно еще увеличить объем кадров до цифр, сопоставимых с количеством объектов другого класса.
2. Похоже, что, на класс 2 Владельцы модель обучилась на одежду, аксессуары (в случае с ребенком на красную ледянку). Если на Владельцах будет другая одежда, то модель примет объект класса Владельцы за класс Люди и выдаст ложную тревогу. Увеличением количества кадров эту проблему не решить. Для распознавания именно Владельцев следует полностью изменить подход - сменить расположение камеры чтобы в кадр всегда попадали лица вошедших и распознавать людей по лицам.

8. Список литературы

Обзор лучших моделей обнаружения объектов 2025 года
Сравнение RT-DETR и YOLO11
Документация на модель YOLO11
Документация на модель RT-DETR