

Dmitry Sidnev
Summer Camp 2021

Object tracking. Overview of modern approaches



План:

1. Метрики
2. Датасеты и челленджи
3. Deep learning подходы

Метрики

- Классические метрики:
 - Mostly Tracked (MT) trajectories: количество ground-truth траекторий, которые корректно сопоставлены минимум на 80% кадров
 - Mostly Lost (ML) trajectories: количество ground truth траекторий, которые корректно сопоставлены менее, чем на 20% кадров
 - Fragments: найденные траектории, которые покрывают не более 80 % ground truth траектории
 - False trajectories: найденные траектории, которые не соответствуют ни одному объекту в ground truth
 - ID switches: количество изменений ID корректно сопоставленной траектории

Метрики

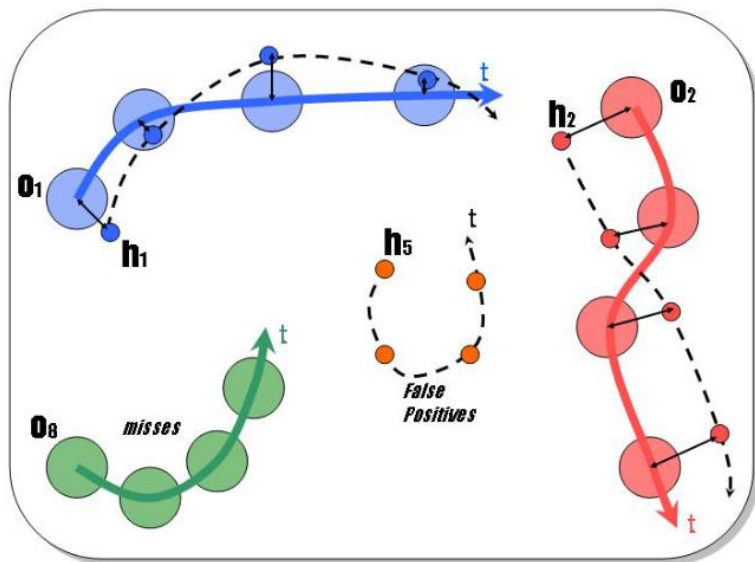
- CLEAR (*Classification of Events, Activities and Relationships*) MOT metrics:

- FP: количество false positives на всем видео
- FN: количество false negatives на всем видео
- Fragm: общее количество фрагментов
- IDSW: общее количество "перескоков" ID

- Multiple object tracking accuracy (MOTA)
$$MOTA = 1 - \frac{(FN + FP + IDSW)}{GT}$$

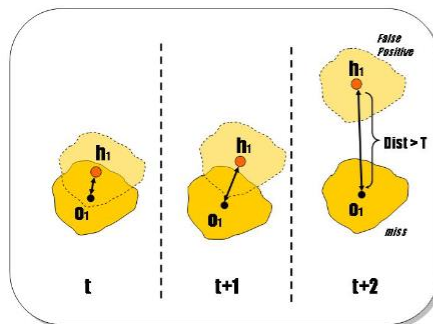
- Multiple object tracking precision (MOTP)
$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}$$

Метрики

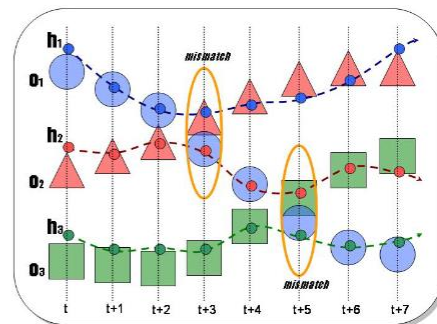


h_1, h_2, h_5 – треки, полученные
нашим алгоритмом

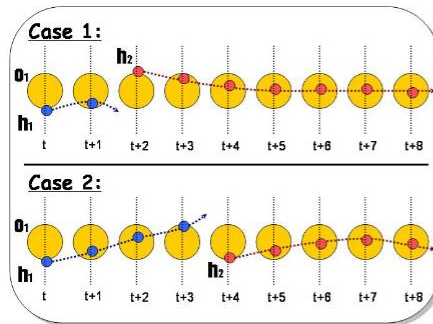
O_1, O_2, O_3 – ground truth



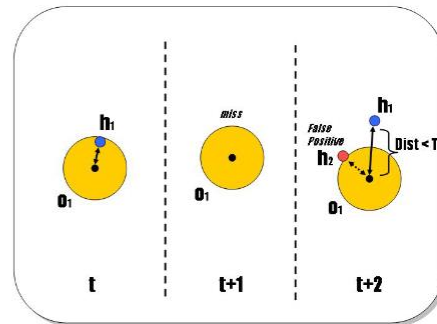
(a)



(b)



(c)



(d)

Метрики

- ID scores: вместо сопоставления ground truth и задетектированных объектов покадрово сопоставление выполняется более глобально и каждая траектория из ground truth соответствует только одной найденной траектории с максимальным количеством корректных кадров

- Identification precision: $IDP = \frac{IDTP}{IDTP + IDFP}$

- Identification recall: $IDR = \frac{IDTP}{IDTP + IDFN}$

- Identification F1: $IDF1 = \frac{2}{\frac{1}{IDP} + \frac{1}{IDR}} = \frac{2IDTP}{2IDTP + IDFP + IDFN}$

Датасеты и челленджи

Самые популярные челленджи для трекинга:

- MOTChallenge: самый известный из всех benchmark'ов для 'multiple object tracking'
 - Содержит тренировочный и тестовый датасеты
 - Содержит так называемые `public` детекшены
 - Классы: пешеходы
 - Датасеты: MOT15, MOT16, MOT17, MOT20
 - <https://motchallenge.net/>
- KITTI: трекинг пешеходов и автомобилей
 - Собран с движущейся по улицам города машины
 - 21 видео для тренировки и 29 для теста
 - Можно загрузить результаты только для пешеходов или только для автомобилей
 - <http://www.cvlibs.net/datasets/kitti/>



Deep learning подходы: online vs offline

- Offline (or Batch) алгоритмы могут использовать информацию с будущих кадров. Такие алгоритмы используют глобальную информацию и, как правило, дают лучше результат
- Online алгоритмы могут использовать только данные с текущего кадра и прошлых (в некоторых прикладных задачах это является необходимым условием, например автономное вождение)
 - Из-за невозможности использовать информацию из будущего данный алгоритм не может исправить ошибки в прошлом

SORT (Simple online and realtime tracking)

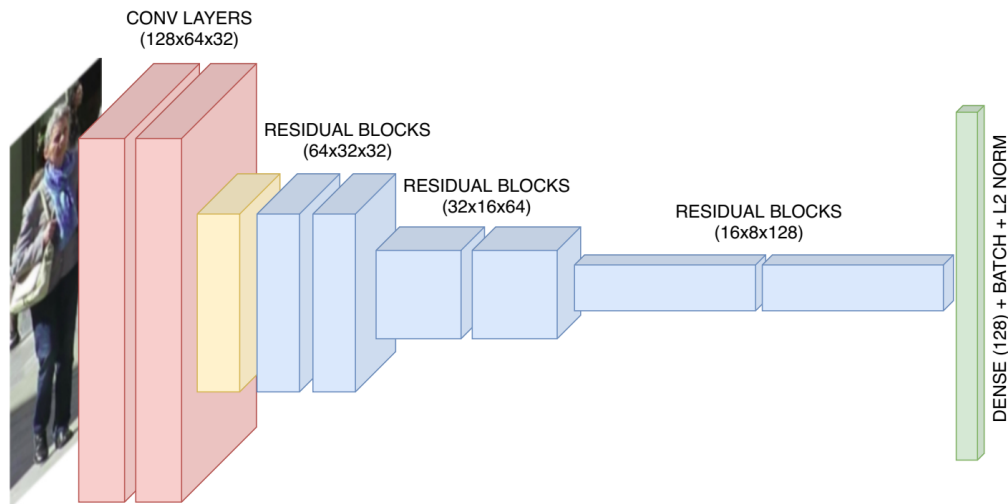
SORT - один из первых MOT трекеров, который использовал нейронные сети для детектирования пешеходов.

- Использование в качестве детектора Faster R-CNN (только это улучшило метрику MOTA на 18.9% на датасете MOT15)
- Прогнозирование движения объекта с помощью Kalman filter
- Связывание детекшенов с помощью Hungarian algorithm
- Использование intersection-over-union (IoU) для вычисления матрицы схожести

<https://arxiv.org/abs/1602.00763>

DeepSORT: evolution of the SORT algorithm

DeepSORT использует дополнительно к SORT алгоритму appearance feature extractor.



- Каждый детекшен с пешеходом пропускается через сверточную сеть с вектором размерностью 128 на выходе, который неким образом описывает внешние признаки пешехода
- В качестве критерия схожести пешеходов вычисляется евклидово или косинусное расстояние между векторами

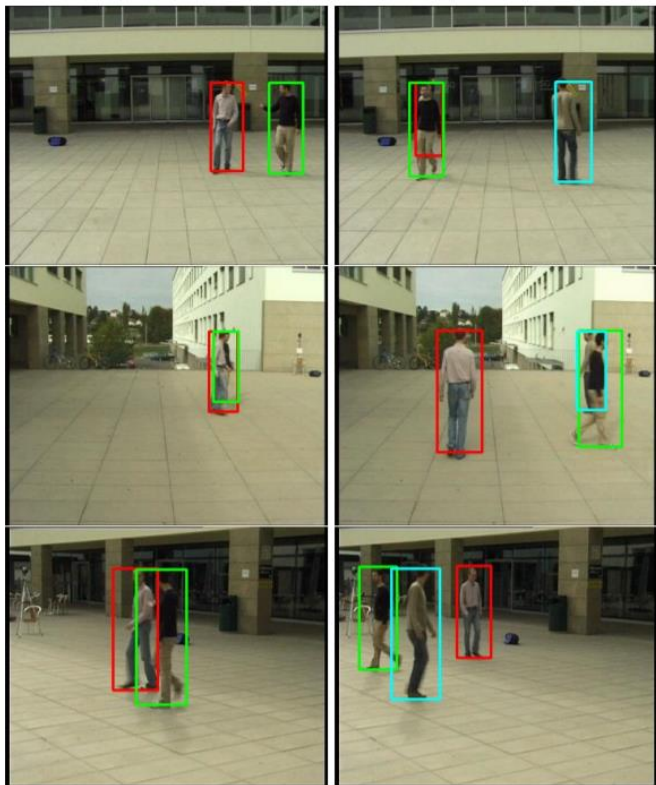
<https://arxiv.org/abs/1703.07402>

Person re-identification based tracking



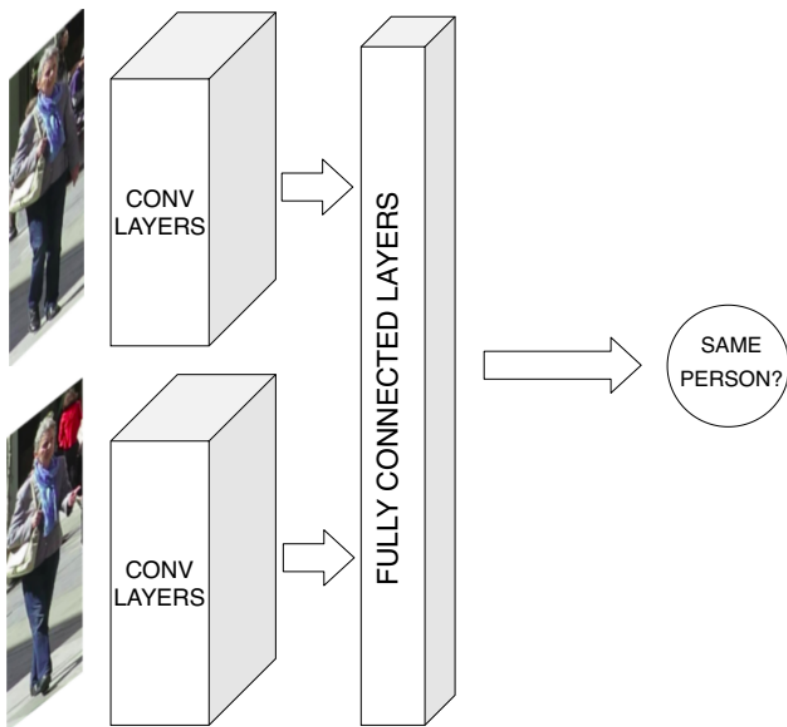
- Person re-identification модели ищут признаки, которые являются общими для одного и того же человека
- Модели тренируются отдельно на специальных датасетах (Market-1501, MSMT17 e.t.c.)
- Качество трекера сильно зависит от качества применяемой Person re-identification модели

Multi camera multi-person tracking



- Более сложная подзадача трекинга - когда мы имеем несколько камер (перекрывающихся или нет) и требуется отследить объект на всех камерах
- Основная сложность заключается в следующем:
 - Разные камеры могут иметь разное качество, разные уровни освещенности и т.д.
 - Люди попадают на камеры под разными ракурсами, что усложняет их идентификацию
- Самый простой способ решения задачи - использование Person re-identification моделей

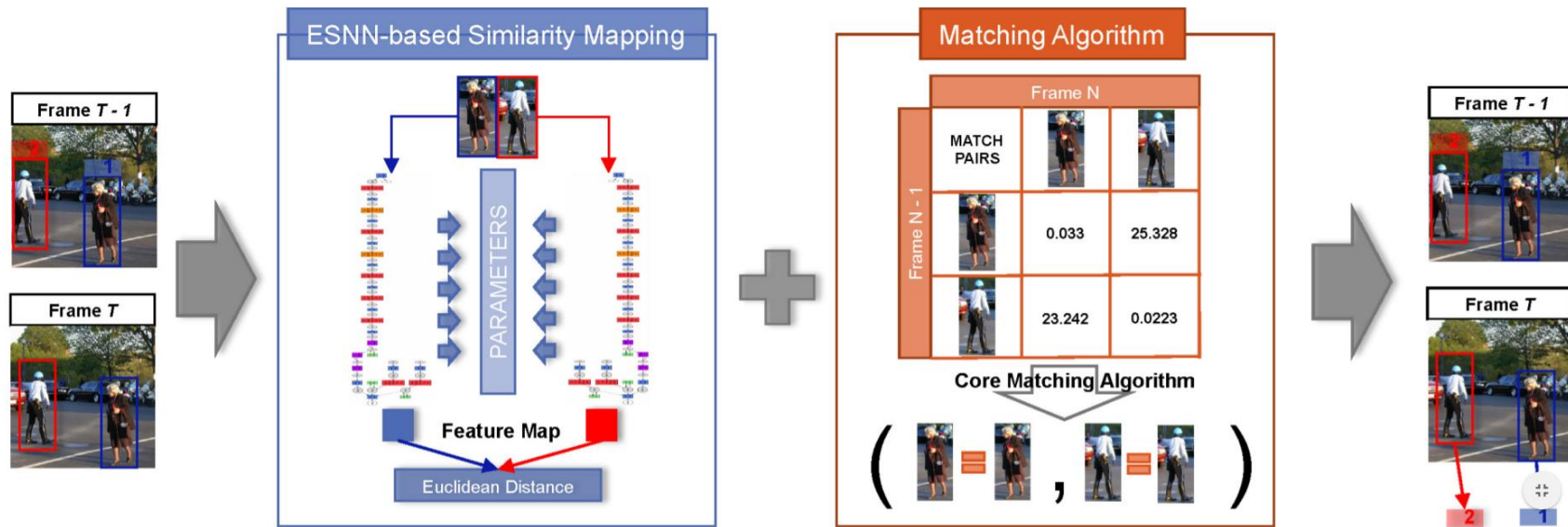
Siamese networks



- Сверточные сети с функцией потерь, которая объединяет информацию с разных изображений для обучения особенностям, которые наилучшим образом отличают объекты друг от друга
- Для использования в трекинге во время инференса функция потерь откидывается и вектор, получаемый с последнего FC слоя может применяться для одиночного детекшена

Примеры использования Siamese networks

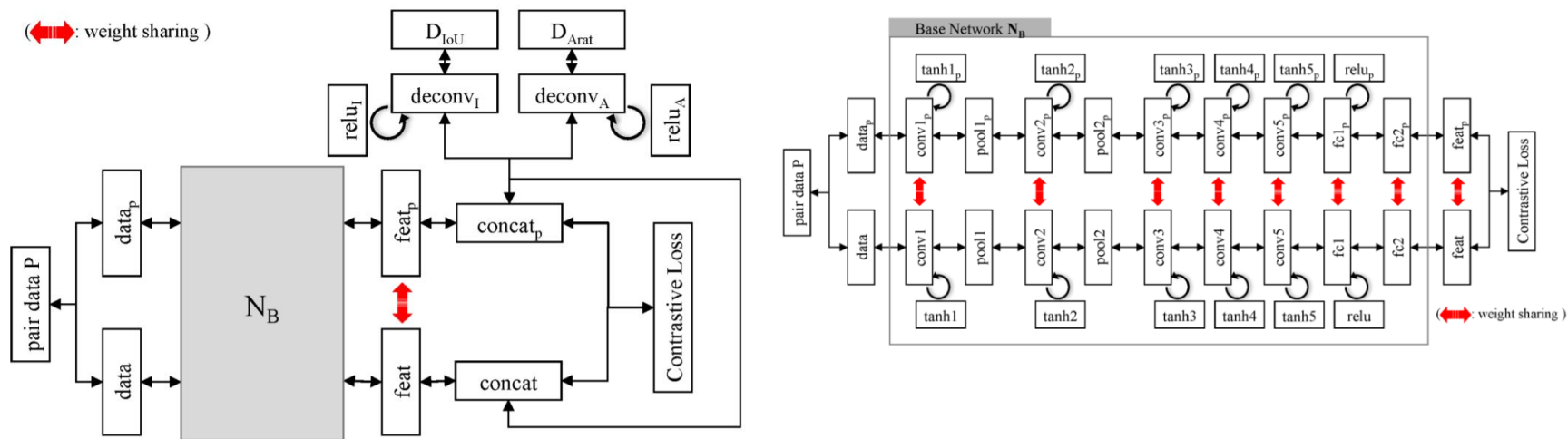
Similarity Mapping with Enhanced Siamese Network for Multi-Object Tracking



<https://arxiv.org/abs/1609.09156>

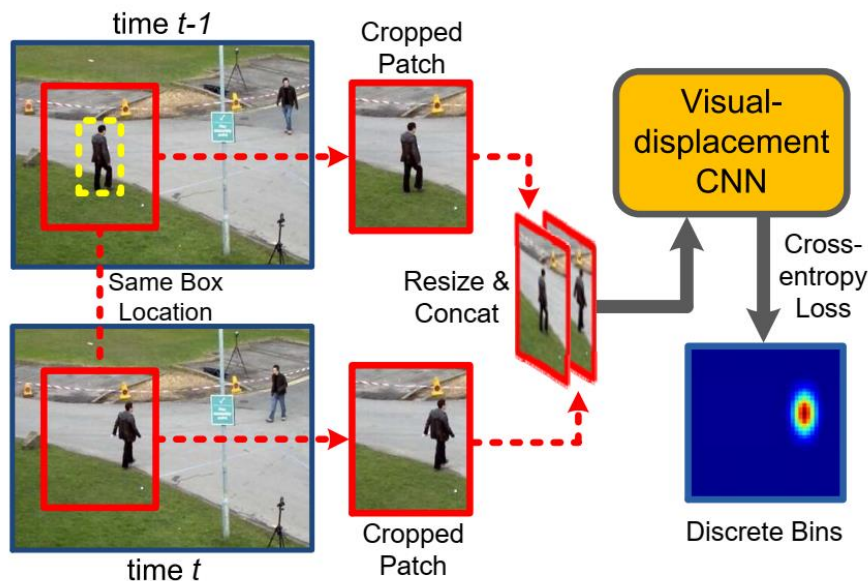
Примеры использования Siamese networks

Similarity Mapping with Enhanced Siamese Network for Multi-Object Tracking



Примеры использования Siamese networks

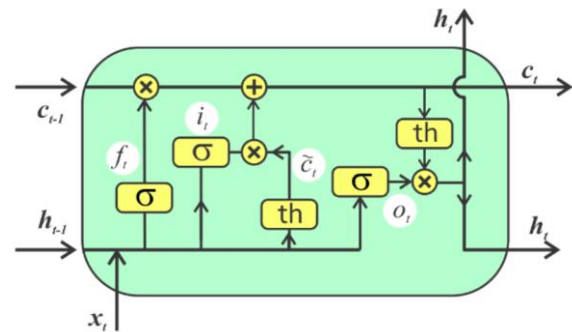
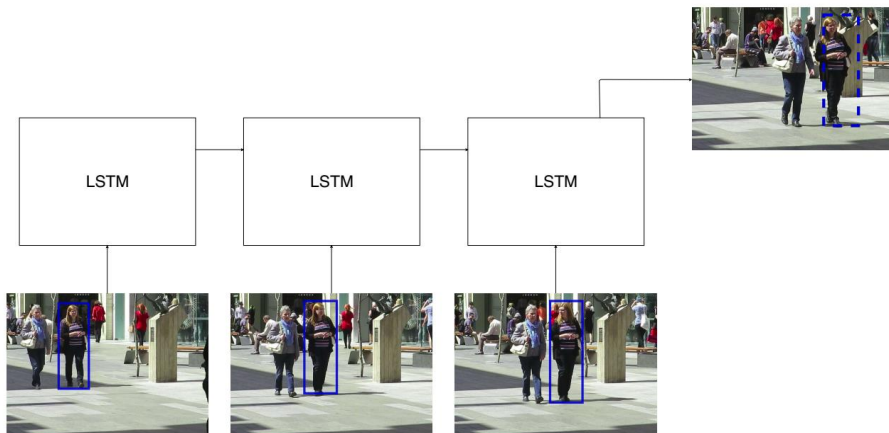
Deep Continuous Conditional Random Fields with Asymmetric Inter-object Constraints for Online Multi-object Tracking



- Две области вырезаются в одном и том же месте на кадре t и $t-1$, конкатенируются и подаются на вход CNN
- Сверточная нейронная сеть на выходе дает оценку визуального перемещения объекта
- Для решения проблемы перекрытия оценивается скорость перемещения для каждой пары объектов
- Hungarian algorithm для сопоставления трека и детекшена

<https://arxiv.org/abs/1806.01183>

RNN based tracking

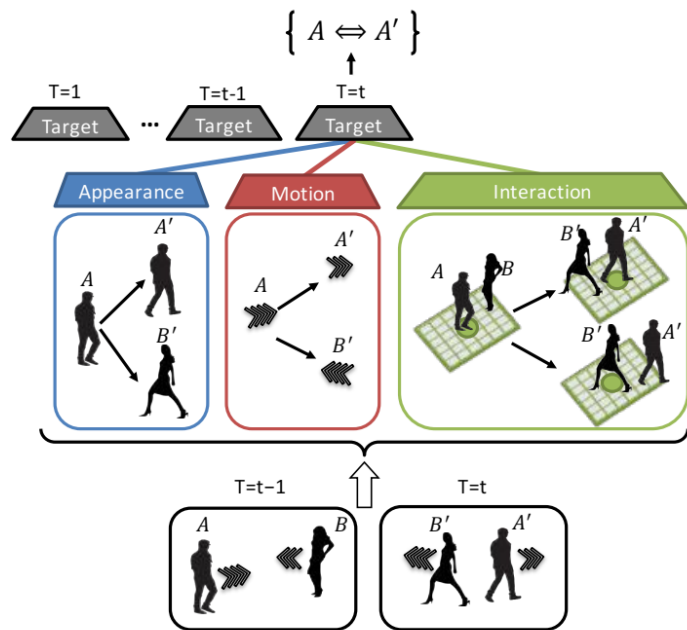


LSTM ячейка

Детекшены кадр за кадром подаются на вход рекуррентной сети, которая на каждом очередном шаге использует состояние и веса предыдущего шага

RNN based tracking

Tracking The Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies

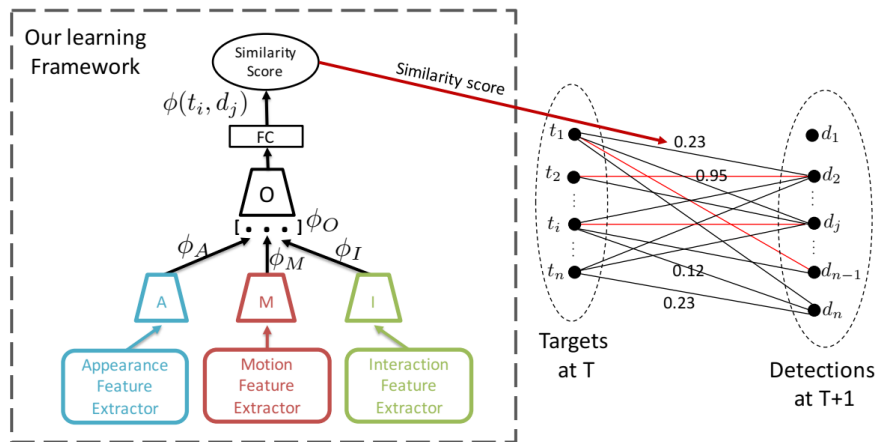


- Алгоритм основан на обработке 3 характеристик объекта:
 - Внешние признаки
 - Перемещение (скорость движения)
 - Взаимодействие с другими объектами (пересечение)
- Каждая из характеристик вычисляется с помощью отдельной рекуррентной сети

<https://arxiv.org/abs/1701.01909>

RNN based tracking

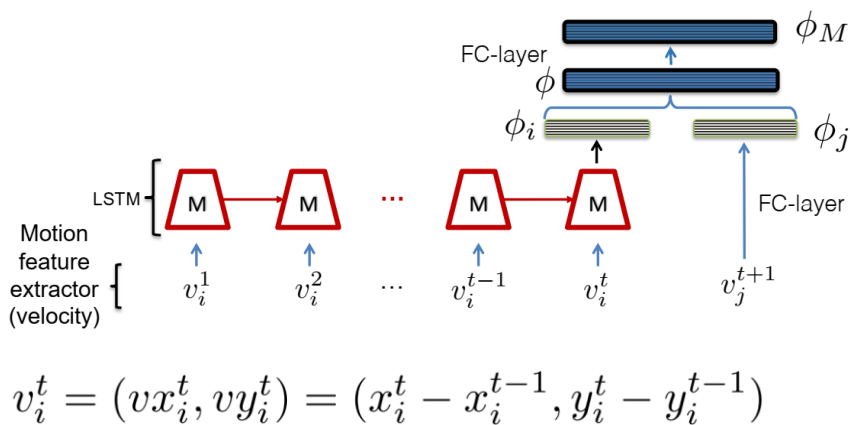
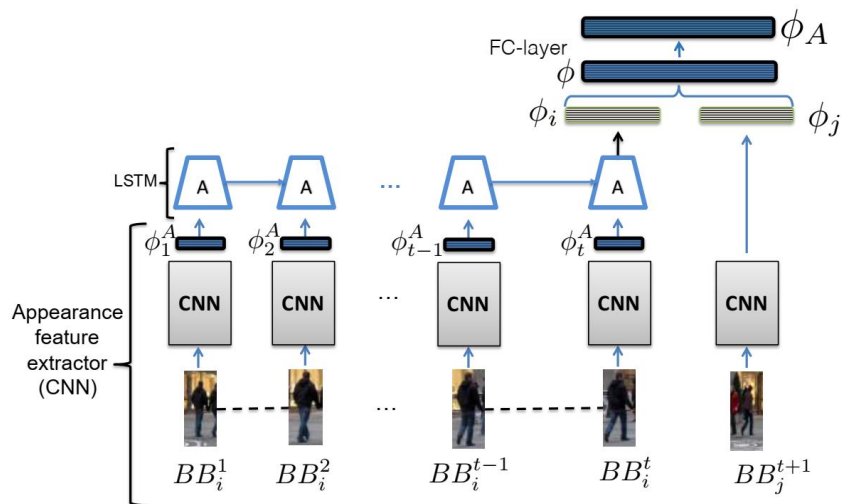
Tracking The Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies



- Все 3 характеристики объединяются и подаются на вход следующей рекуррентной нейронной сети
- Затем на выходе для каждой пары 'треклет – детекшен' получаем вектор для оценки схожести

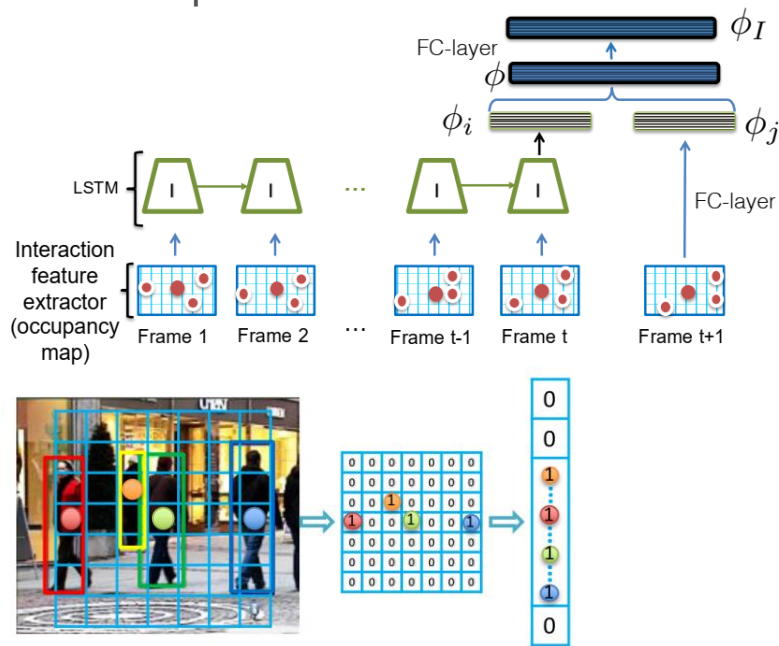
RNN based tracking

Tracking The Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies



RNN based tracking

Tracking The Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies



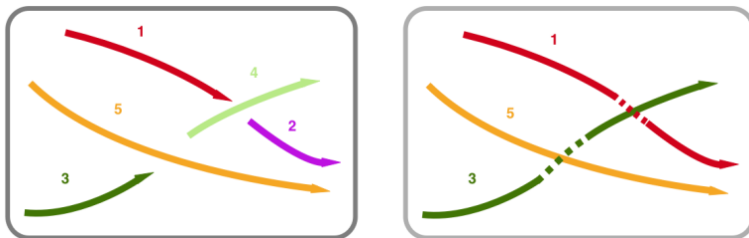
Accupancy map

Ключевые моменты:

- В качестве CNN для извлечения внешних признаков использовалась сеть VGG16
- CNN для извлечения внешних признаков тренировалась, как person re-identification модель
- Количество кадров, необходимое для LSTM зависит от видео и того, как долго объекты перекрываются на видео (для MOT16 достаточно 6 кадров)
- Алгоритм показывает хорошие результаты на MOT16

RNN based tracking

Occlusion handling in tracking multiple people using RNN



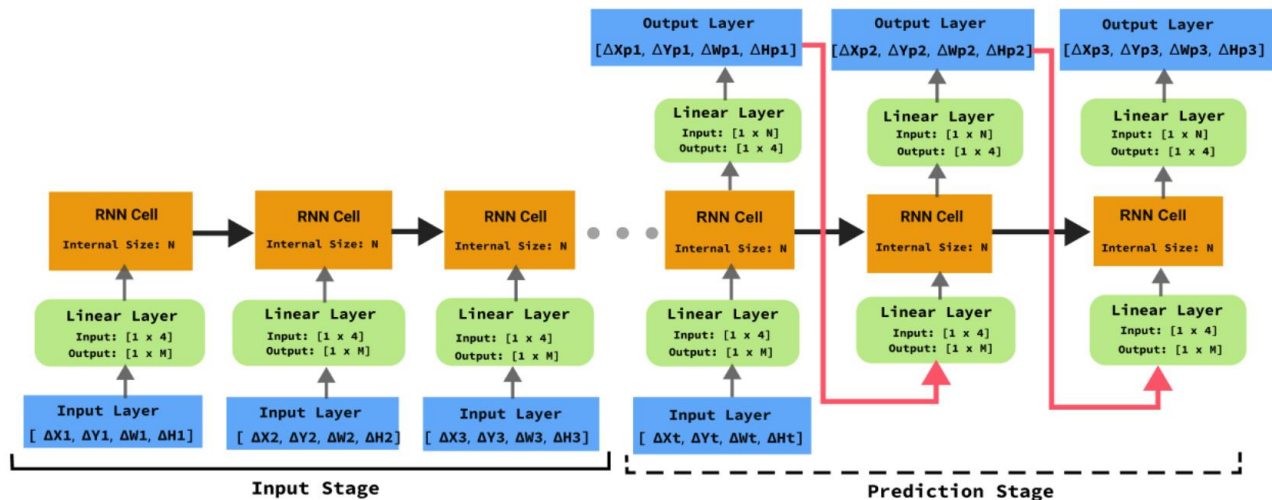
Ключевые моменты:

- Основная решаемая проблема - перекрытия объектов или незадетекченные объекты, когда трек прерывается
- При помощи рекуррентной нейронной сети предсказывают баундинг боксы в "будущем" или "прошлом"
- Треклеты, разнесенные во времени и удовлетворяющие заданным критериям, склеиваются в один непрерывный

<https://ieeexplore.ieee.org/document/8451140>

RNN based tracking

Occlusion handling in tracking multiple people using RNN



Алгоритм: на вход сети подаются величины $\Delta X, \Delta Y, \Delta W, \Delta H$, описывающие изменение координат и размера баундинг бокса между двумя соседними кадрами, до тех пор пока треклет не прервется. Затем начинается фаза предсказания положения и размера баундинг бокса

RNN based tracking

Occlusion handling in tracking multiple people using RNN

$$overlap_1 + overlap_2 \geq stitch_thr$$

(1) Критерий слияния двух треклетов основан на intersection over union (IoU) между:

$$overlap_1 = \frac{T_1(t) \cap P_2(t)}{T_1(t) \cup P_2(t)},$$

(2)

- **T1** - треклет, который закончился на кадре t
- **T2** - треклет, который является потенциальным кандидатом для слияния на кадре $t + \Delta t$

$$overlap_2 = \frac{P_1(t + \Delta t) \cap T_2(t + \Delta t)}{P_1(t + \Delta t) \cup T_2(t + \Delta t)}.$$

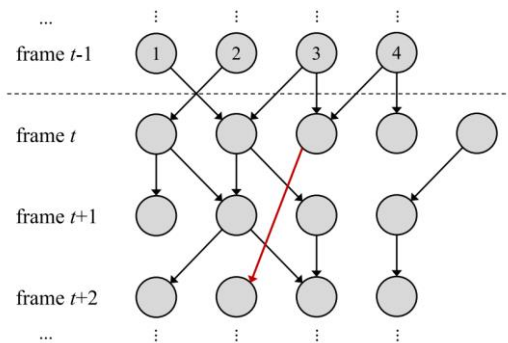
(3)

- **P1** - предсказанный баундинг бокс на кадре $t + \Delta t$ для T1 (в будущем)
- **P2** - предсказанный баундинг бокс на кадре t для T2 (т.е. в прошлом)

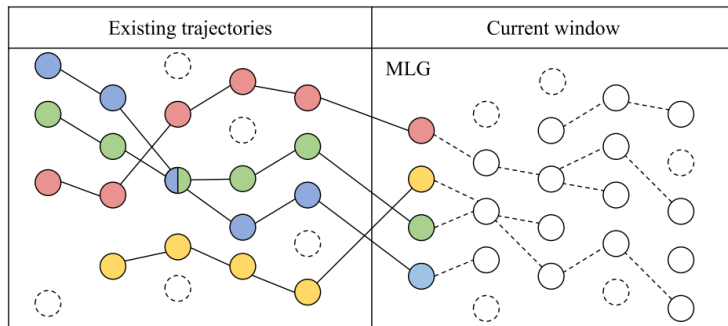
Сумма IoU между (T1, P2) и (T2, P1) должна удовлетворять условию (1)

RNN based tracking

Multiplex Labeling Graph for Near-Online Tracking in Crowded Scenes



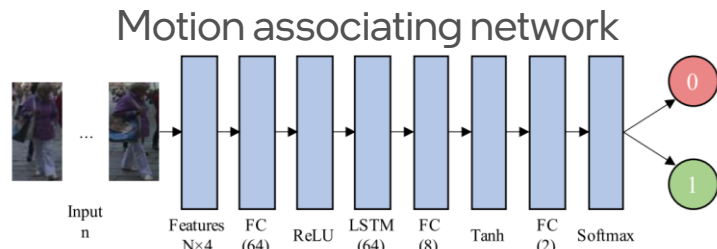
- Основная обозначенная проблема - перекрытие объектов
- По видео проходит скользящее окно, которое охватывает N кадров
- Треклеты на первом кадре в данном окне имеют свои ID, строится однонаправленный граф для всех объектов, существующих в данном окне
- Каждая вершина графа (т.е. объект) может иметь больше одного ребра, т.е. разные треклеты могут иметь общие детекшены



<https://ieeexplore.ieee.org/document/9098857>

RNN based tracking

Multiplex Labeling Graph for Near-Online Tracking in Crowded Scenes



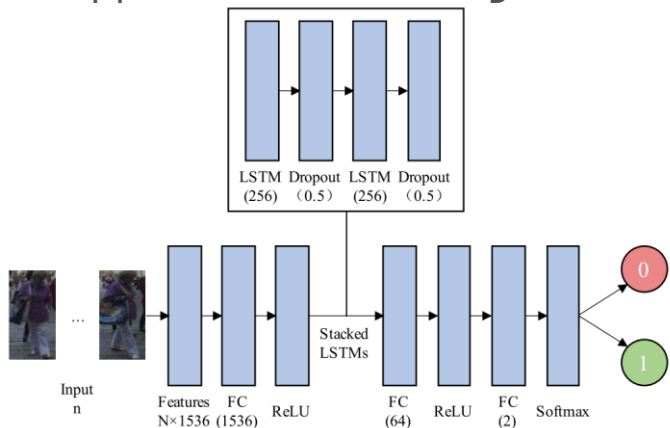
$$S = S_m + S_a$$

S - критерий соединения вершин графа

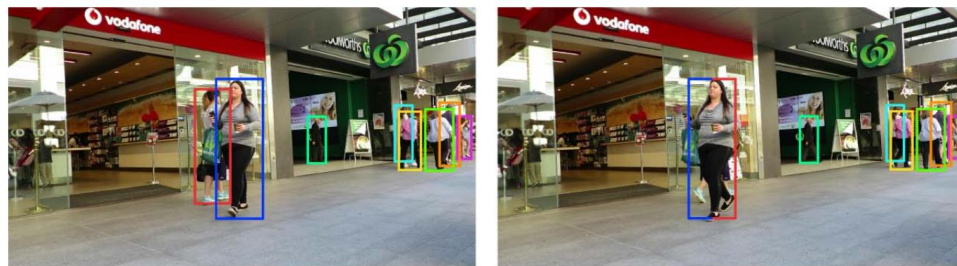
S_m – motion feature

S_a – apperance feature

Appearance associating network



Пример работы трекера



RNN based tracking

Multiplex Labeling Graph for Near-Online Tracking in Crowded Scenes















RESULTS ON MOT CHALLENGE 2017 TEST(2020.2)

| Method | MOTA↑ | IDF1↑ | MT↑ | ML↓ | FP↓ | FN↓ | FP+FN↓ | IDS↓ | FM↓ | detector | type |
|-------------------|-------------|-------------|--------------|--------------|---------------|----------------|----------------|--------------|--------------|----------|--------|
| MLT (ours) | 54.8 | 62.9 | 24.2% | 37.9% | 19,118 | 234,303 | 253,421 | 1,077 | 2,188 | public | batch |
| LSST17[57] | 54.7 | 62.3 | 20.4% | 40.1% | 26,091 | 228,434 | 254,525 | 1,243 | 3,726 | public | batch |
| Tracktor17[47] | 53.5 | 52.3 | 19.5% | 36.6% | 12,201 | 248,047 | 260,248 | 2,072 | 4,611 | public | batch |
| JBNOT[58] | 52.6 | 50.8 | 19.7% | 35.8% | 31,572 | 232,659 | 264,231 | 3,050 | 3,792 | public | batch |
| eTC17[59] | 51.9 | 58.1 | 23.1% | 35.5% | 36,164 | 232,783 | 268,947 | 2,288 | 3,071 | public | batch |
| eHAF[29] | 51.8 | 54.7 | 23.4% | 37.9% | 33,212 | 236,772 | 269,984 | 1,834 | 2,739 | public | batch |
| AFN17[51] | 51.5 | 46.9 | 20.6% | 35.5% | 22,391 | 248,420 | 270,811 | 2,593 | 4,308 | public | batch |
| FWT[60] | 51.3 | 47.6 | 21.4% | 35.2% | 24,101 | 247,921 | 272,022 | 2,648 | 4,279 | public | batch |
| NOTA[48] | 51.3 | 54.5 | 17.1% | 35.4% | 20,148 | 252,531 | 272,679 | 2,285 | 5,798 | public | batch |
| LSST17O[57] | 52.7 | 57.9 | 17.9% | 36.6% | 22,512 | 241,936 | 264,448 | 2,167 | 7,443 | public | online |
| FAMNet[61] | 52.0 | 48.7 | 19.1% | 33.4% | 14,138 | 253,616 | 267,754 | 3,072 | 5,318 | public | online |

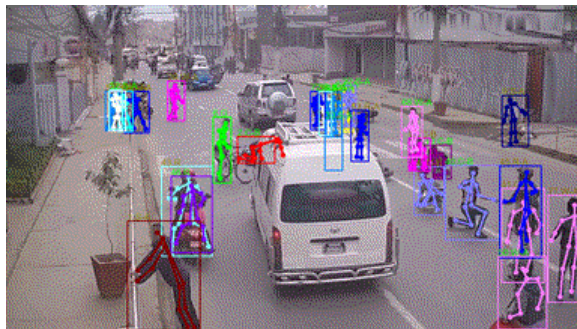
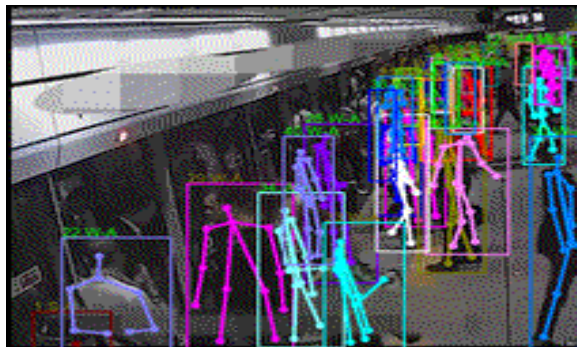
RNN based tracking

Multiplex Labeling Graph for Near-Online Tracking in Crowded Scenes

MOT17 leaderboard (public detections)

| Tracker | ↑MOTA | IDF1 | MOTP | MT | ML | FP | FN | Recall | Precision | FAF | ID Sw. | Frag | Hz |
|--|-------------------|------------------|-------------|---------------------|-------------------|--------|----------------|-------------|-----------|-----|--------------|---------------------|------|
| MLT 1.   | 75.3 ±12.0 | 75.5 ±5.9 | 81.7 | 1,161 (49.3) | 459 (19.5) | 27,879 | 109,836 | 80.5 | 94.2 | 1.6 | 1,719 (21.3) | 1,737 (21.6) | 5.9 |
| Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Ke, Z. Xiong. Multiplex Labeling Graph for Near Online Tracking in Crowded Scenes . In IEEE Internet of Things Journal, 2020. | | | | | | | | | | | | | |
| TraJE 2.   | 67.8 ±15.4 | 61.4 ±10.8 | 78.3 | 848 (36.0) | 578 (24.5) | 20,982 | 157,468 | 72.1 | 95.1 | 1.2 | 3,475 (48.2) | 5,668 (78.6) | 1.4 |
| MAT 3.   | 67.1 ±13.1 | 69.2 ±10.0 | 80.8 | 917 (38.9) | 622 (26.4) | 22,756 | 161,547 | 71.4 | 94.7 | 1.3 | 1,279 (17.9) | 2,037 (28.5) | 11.5 |
| MAT: Motion-Aware Multi-Object Tracking | | | | | | | | | | | | | |
| RCNN_T 4.  | 63.9 ±14.2 | 66.1 ±10.5 | 79.4 | 795 (33.8) | 655 (27.8) | 22,565 | 179,568 | 68.2 | 94.5 | 1.3 | 1,774 (26.0) | 4,182 (61.3) | 59.2 |
| SSAT 5.   | 62.0 ±16.2 | 62.6 ±12.3 | 78.9 | 650 (27.6) | 748 (31.8) | 14,970 | 197,670 | 65.0 | 96.1 | 0.8 | 1,850 (28.5) | 4,911 (75.6) | 3.2 |
| UnsupTrack 6.   | 61.7 ±16.0 | 58.1 ±11.1 | 78.3 | 640 (27.2) | 762 (32.4) | 16,872 | 197,632 | 65.0 | 95.6 | 1.0 | 1,864 (28.7) | 4,213 (64.8) | 2.0 |
| S. Karthik, A. Prabhu, V. Gandhi. Simple Unsupervised Multi-Object Tracking . In Arxiv, 2020. | | | | | | | | | | | | | |
| CTTrackPub 7.   | 61.5 ±16.1 | 59.6 ±12.2 | 78.9 | 621 (26.4) | 752 (31.9) | 14,076 | 200,672 | 64.4 | 96.3 | 0.8 | 2,583 (40.1) | 4,965 (77.1) | 17.0 |
| X. Zhou, V. Koltun, P. Kráhenbühl. Tracking Objects as Points . In ECCV, 2020. | | | | | | | | | | | | | |
| Lif_T 8.  | 60.5 ±13.0 | 65.6 ±8.6 | 78.3 | 637 (27.0) | 791 (33.6) | 14,966 | 206,619 | 63.4 | 96.0 | 0.8 | 1,189 (18.8) | 3,476 (54.8) | 0.5 |
| A. Hornakova, R. Henschel, B. Rosenhahn, P. Swoboda. Lifted Disjoint Paths with Application in Multiple Object Tracking . In ICML, 2020. | | | | | | | | | | | | | |

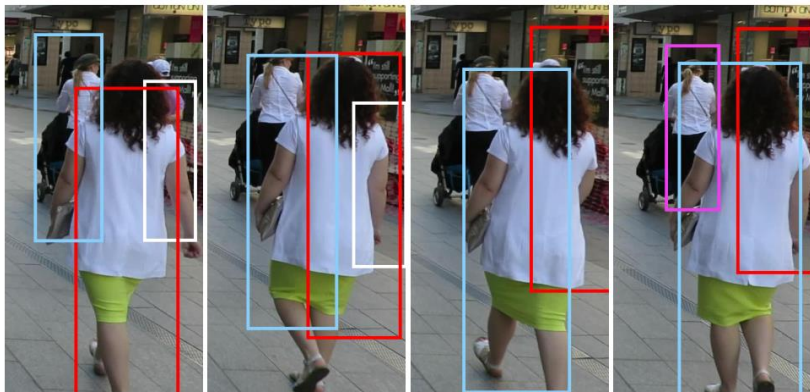
Pose tracking



Существует еще один челлендж - **pose tracking** (относительно новый), где необходимо сопровождать объект не по баундинг боксу, а по скелетону, полученному с помощью pose estimation сети: <http://humanevents.org/>

- Помимо скелетонов так же доступна привычная аннотация с баундинг боксами
- Для участия доступны несколько треков, в том числе и Multi-person Motion Tracking
- Возможно трекинг на основе скелетонов может дать более точный результат, учитывая более точную информацию о положении человека на кадре относительно баундинг бокса

Заключение



Пример перекрытия объектов с последующим ID switch'ем

- Самая распространенная проблема в задаче трекинга - пересечение объектов, когда один из них частично или полностью перекрывает другой
- Сильное влияние детектора на качество трекинга
- Feature extractors необходимо тренировать отдельно, что требует дополнительных данных и времени для тренировки
- Более качественные решения более требовательны к ресурсам
- Как конкретное решение будет работать с другими камерами, в других условиях и .т.д.?

The Intel logo is centered on a solid blue background. It features the word "intel" in a white, lowercase, sans-serif font. A small blue square is positioned above the letter "i". To the right of the word "intel" is a registered trademark symbol (®).

intel®