# Оценивание параметров в задаче слежения за множеством объектов в видеопотоке

Григорьев Алексей Дмитриевич

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Научный руководитель: к.ф.-м.н. А.Н. Гнеушев

Москва, 2020

# Цель исследования

## Задача

Разработать модель сопровождения объектов по видео последовательности кадров, основанную на ре-идентификации объектов.

## Проблема

Существующие решения являются либо ресурсоемкими, либо неустойчивыми к большой плотности объектов, что приводит к срыву слежения, перескоку траекторий, накоплению ошибок с течением временем.

## Решение

Использовать подход отбора объектов для ре-идентификации на основе оценки качества. Такое решение увеличит вычислительную эффективность метода, а также уменьшит число ошибок ре-идентификации.

# Существующие решения

**Задача слежения за множеством объектов**

- *A. Bewley, G. Zongyuan, F. Ramos, B. Upcroft* Simple online and realtime tracking // ICIP, 2016, P. 3464–3468.
- *N. Wojke, A. Bewley, and D. Paulus* Simple online and realtime tracking with a deep association metric // ICIP, 2017, P. 3645–3649.

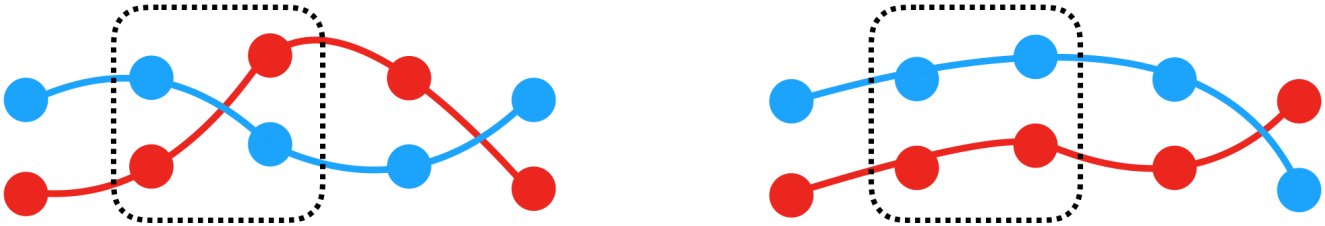
**Задача оценки качества**

- *L. Best-Rowden, A. K. Jain. Learning Face* Image Quality From Human Assessments // IEEE Transactions on Information Forensics and Security, 2018.

# Разрыв траектории. Ре-идентификация

## Существующая проблема

Некорректные переключения между объектами и разрывы траекторий при их продлении в сценариях с пересекающимися и временно пропадающими из кадра объектами, имеющими сложную модель движения.



Истинные траектории (слева), ложные переключения (справа)

## Возможное решение

Ре-идентификация – задача принятия решения о наличии объекта, присутствующего на предоставленным изображении, на множестве других имеющихся изображений.

## Обозначения

$Y_t = \{\mathbf{y}_{j,1}, \ldots, \mathbf{y}_{j,t-1}, |j = 1, \ldots, N\}$ – измерения всех $N$ объектов на предыдущих кадрах, $X_t = \{\mathbf{x}_{j,t}, |j = 1, \ldots, N\}$ – скрытые характеристики (состояния) всех $N$ объектов на кадре $t$, подлежащие оцениванию, $Z_t = \{\mathbf{z}_{1,t}, \ldots, \mathbf{z}_{M_t,t}\}$ – измеренные характеристики не идентифицированных объектов на кадре t с помощью детектора объектов.

## Допущения

- $\mathbf{x}_{j,t}$ зависит только от $\mathbf{x}_{j,t-1}$; $\mathbf{y}_{j,t}$ определяются только $\mathbf{x}_{j,t}$;
- процесс изменения характеристик $\mathbf{x}_{j,t}$ каждого объекта описывается моделью ЛДС, нормальность распределений:

$$p(\mathbf{x}_{j,t}|\mathbf{x}_{j,t-1}) = N(\mathbf{A}_j\mathbf{x}_{j,t-1}, \mathbf{\Gamma}_j),$$
$$p(\mathbf{y}_{j,t}|\mathbf{x}_{j,t}) = N(\mathbf{B}_j\mathbf{x}_{j,t}, \mathbf{\Sigma}_j),$$
$$p(\mathbf{x}_{j,1}) = N(\boldsymbol{\mu}_{j,0}, \mathbf{\Gamma}_{j,0}).$$

- объекты не взаимодействуют и двигаются независимо, одно обнаружение на кадре связано только с одним объектом;
- $\mathbf{x} = (u, v, s, r, \dot{u}, \dot{v}, \dot{s})^T$, где $u$ и $v$ – координаты, $s$ и $r$ – площадь и соотношение сторон прямоугольника локализации, $\dot{u}, \dot{v}, \dot{s}$ - скорости.

## Исходная задача – максимизация совместной апостериорной вероятности

$$\max_{X_t} p(X_t|Y_t, Z_t)$$

## Факторизованная апостериорная вероятность

$$p(X_t|Y_t, Z_t) = \prod_{j=1}^{N} \left\{ \sum_{i=1}^{M_t} a_{i,j}\, p(\mathbf{x}_{j,t}|\mathbf{y}_{j,1}, \ldots, \mathbf{y}_{j,t-1}, \mathbf{z}_{i,t}) \right\}$$

при условиях: $a_{i,j} \in \{0, 1\}$, $\sum_{i=1}^{M_t} a_{i,j} = 1$, $\sum_{j=1}^{N} a_{i,j} = 1$.

## Подзадачи

$$\max_{a_{i,j}} \log p(\hat{X}_t|Y_t, Z_t) = \max_{a_{i,j}} \sum_{j=1}^{N} \sum_{i=1}^{M_t} a_{i,j}\, \log L_{\hat{\mathbf{x}}_{j,t}}(\mathbf{y}_{j,1}, \ldots, \mathbf{y}_{j,t-1}, \mathbf{z}_{i,t})$$

$$\max_{X_t} \log p(X_t|Y_t, Z_t) = \sum_{j=1}^{N} \max_{\mathbf{x}_{j,t}} \log p(\mathbf{x}_{j,t}|\mathbf{y}_{j,1}, \ldots, \mathbf{y}_{j,t-1}, \mathbf{y}_{i,t}), \mathbf{y}_{j,t} = \sum_{i=1}^{M_t} a_{i,j}\mathbf{z}_{i,t}$$

## Шаг 1. Прогноз.

$$\hat{\mathbf{x}}_{j,t} = \arg\max_{\mathbf{x}_{j,t}} \log p(\mathbf{x}_{j,t}|\mathbf{y}_{j,1},\ldots,\mathbf{y}_{j,t-1}) = \arg\max_{\mathbf{x}_{j,t}} \log N(\mathbf{x}_{j,t}|\hat{\mathbf{x}}_{j,t},\hat{\mathbf{V}}_{j,t})$$

$$\hat{\mathbf{x}}_{j,t} = \mathbf{A}_j\boldsymbol{\mu}_{j,t-1}, \hat{\mathbf{V}}_{j,t} = \boldsymbol{\Gamma}_j + \mathbf{A}_j\hat{\hat{\mathbf{V}}}_{j,t-1}\mathbf{A}_j^T.$$

## Шаг 2. Задача о назначениях между $Z_t$ и $\{\mathbf{y}_{j,t}|j=1,\ldots,N\}$

$$\min_{a_{i,j}} \sum_{j=1}^{N}\sum_{i=1}^{M_t} a_{i,j}\,\mathbf{C}_{i,j},\ \mathbf{y}_{j,t} = \sum_{i=1}^{M_t} a_{i,j}\mathbf{z}_{i,t}$$

где $\mathbf{C}_{i,j} = -\log L_{\hat{\mathbf{x}}_{j,t}}(\mathbf{y}_{j,1},\ldots,\mathbf{y}_{j,t-1},\mathbf{z}_{i,t})$, $L_{\hat{\mathbf{x}}_{j,t}}$ – функция правдоподобия;

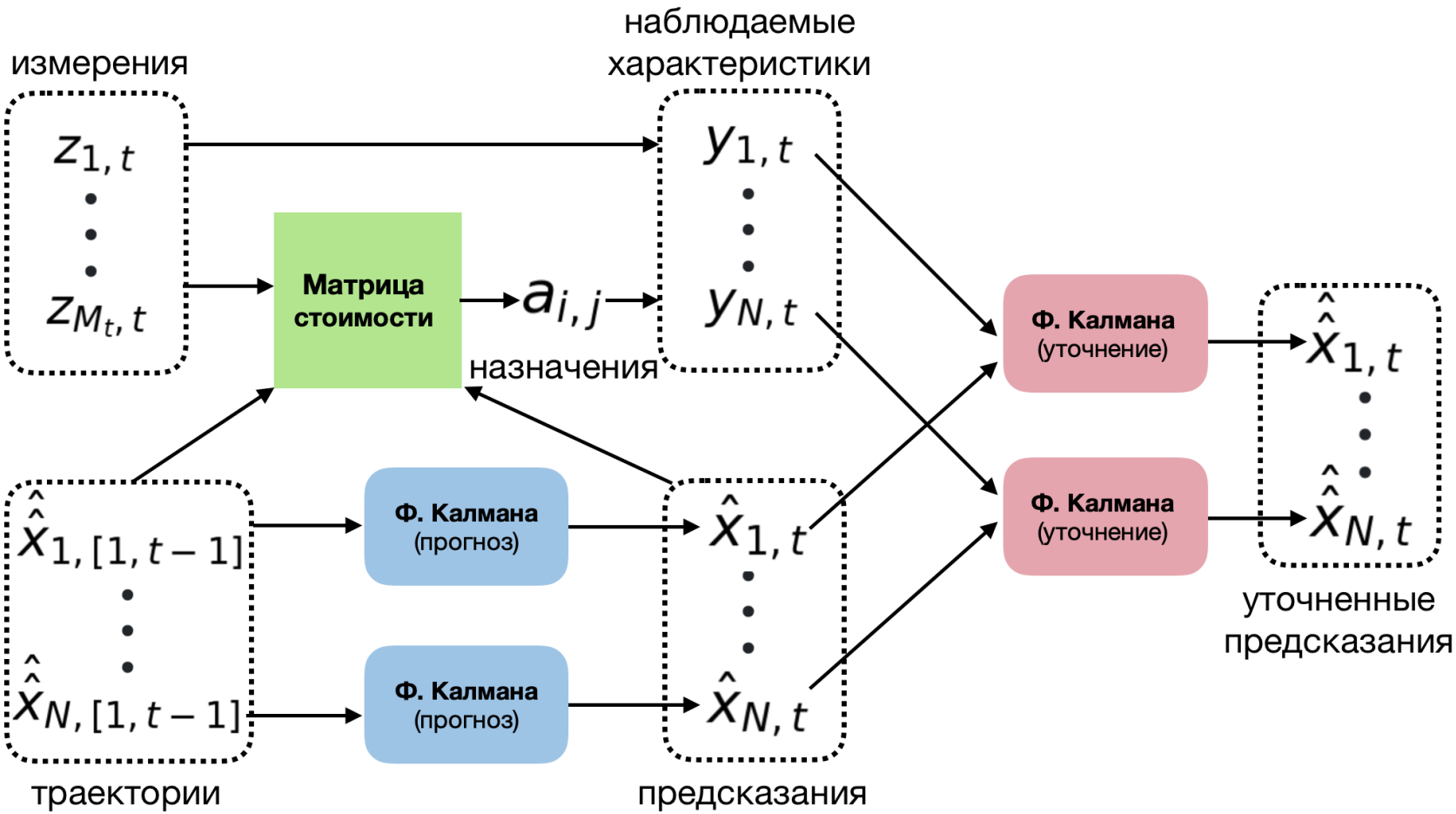
## Шаг 3. Коррекция.

$$\hat{\hat{\mathbf{x}}}_{j,t} = \arg\max_{\mathbf{x}_{j,t}} \log p(\mathbf{x}_{j,t}|\mathbf{y}_{j,1},\ldots,\mathbf{y}_{j,t-1},\mathbf{y}_{j,t}) = \arg\max_{\mathbf{x}_{j,t}} \log N(\mathbf{x}_{j,t}|\hat{\hat{\mathbf{x}}}_{j,t},\hat{\hat{\mathbf{V}}}_{j,t})$$

$$\hat{\hat{\mathbf{x}}}_{j,t} = \hat{\mathbf{x}}_{j,t} + \mathbf{K}_{j,t}(\mathbf{y}_{j,t} - \mathbf{B}_j\hat{\mathbf{x}}_{j,t}), \hat{\hat{\mathbf{V}}}_{j,t} = (\mathbf{I} - \mathbf{K}_{j,t}\mathbf{B}_j)\hat{\mathbf{V}}_{j,t},$$

$$\mathbf{K}_{j,t} = \hat{\mathbf{V}}_{j,t}\mathbf{B}_j^T(\mathbf{B}_j\hat{\mathbf{V}}_{j,t}\mathbf{B}_j^T + \boldsymbol{\Sigma}_j)^{-1}.$$

# Схема решения задачи слежения на кадре $t$




Алгоритм слежения за множеством объектов для кадра $t$

# Дескрипторы объектов для ре-идентификации

- Определяется отображение $h(\cdot|\boldsymbol{\theta})$ пространства изображений, содержащих объект исследуемого класса, в пространство дескрипторов — векторное пространство фиксированной размерности, на котором определена операция скалярного произведения.
- Отображение $h(\cdot|\boldsymbol{\theta})$ здается сверточной нейронной сетью архитектуры ResNet18 с $L_2$-нормализацией выхода (Re-ID сеть).
- Нейросеть обучается в задаче классификации объектов исследуемого класса с Cosine Softmax Cross-Entropy функцией потерь:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\kappa \cdot \tilde{\mathbf{W}}_{c_i}\mathbf{d}_i\right)}{\sum\limits_{j=1}^{N}\exp\left(\kappa \cdot \tilde{\mathbf{W}}_{j}\mathbf{d}_j\right)}, \quad \tilde{\mathbf{W}} = \frac{\mathbf{W}}{\|\mathbf{W}\|_2}, \quad \|\mathbf{d}_i\|_2 = 1,$$

где $\mathbf{W}$ — матрица весов, $\mathbf{d}_i = h(\mathbf{I}_i|\boldsymbol{\theta})$ — декриптор, $c_i$ — метка класса, $\kappa$ — параметр масштабирования.

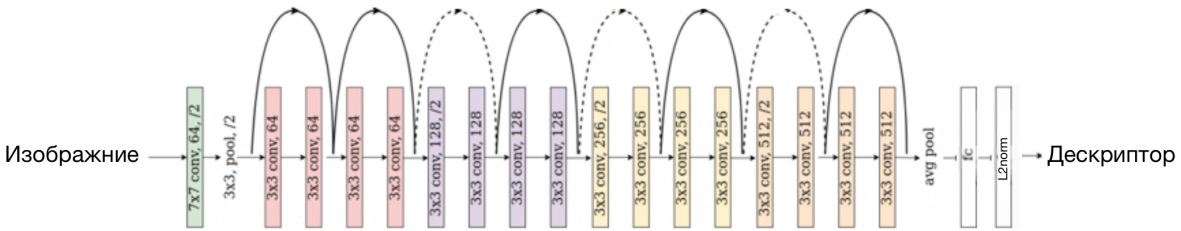



Схема слоев архитектуры Resnet18 с $L_2$-нормализацией выхода

## Связь с задачей сопровождения

После обнаружения и локализации объектов на кадре требуется их сопоставить с ранее найденными траекториями, чтобы для каждой из траектории произвести уточнение прогноза.

## Матрица стоимости

$$\mathbf{C}_{i,j} - \log L_{\hat{\mathbf{x}}_{j,t}}(\mathbf{y}_{j,1}, \ldots, \mathbf{y}_{j,t-1}, \mathbf{z}_{i,t}) = -\log N(\mathbf{z}_{i,t}|\mathbf{B}_j\hat{\mathbf{x}}_{j,t}, \boldsymbol{\Sigma}_j) =$$
$$= \frac{1}{2}(\mathbf{z}_{i,t} - \mathbf{B}_j\hat{\mathbf{x}}_{j,t})^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{z}_{i,t} - \mathbf{B}_j\hat{\mathbf{x}}_{j,t}) + \frac{m}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}_j|,$$

где $m$ размерность вектора $\mathbf{x}$;

## Поправка к матрице стоимости

$\mathbf{I}_i$ — изображение наблюдения $i$ на текущем кадре,
$\mathbf{I}_{j,1}, \ldots, \mathbf{I}_{j,\tau}$ — изображения в траектории объекта $j$.
Выделяются дескрипторы $\mathbf{d}_i = h(\mathbf{I}_i|\boldsymbol{\theta})$; $\mathbf{d}_{j,t} = h(\mathbf{I}_{j,t}|\boldsymbol{\theta})$, $t \in \overline{1,\tau}$.

$$\tilde{\mathbf{C}}_{i,j} = \mathbf{C}_{i,j} + r\left[1 - \max\left\{\frac{\langle\mathbf{d}_i, \mathbf{d}_{j,1}\rangle}{\|\mathbf{d}_i\|\,\|\mathbf{d}_{j,1}\|}, \ldots, \frac{\langle\mathbf{d}_i, \mathbf{d}_{j,\tau}\rangle}{\|\mathbf{d}_i\|\,\|\mathbf{d}_{j,\tau}\|}\right\}\right]^2,$$

где $r$ - масштабный множитель.

# Отбор кандидатов для Re-ID. Показатель "качества".

## Качество в биометрии

Качество объекта — мера полезности объекта для задачи распознавания.

## Отбор кандидатов для ре-идентификации

- $\mathbf{I}_1, \ldots \mathbf{I}_\tau$ — изображения в траектории, функция оценки качества $q : \mathbf{I}_t \mapsto q_t$, $q_t \in [0, 1]$ — показатель "качества".
- Производится сортировка элементов в порядке убывания показателя "качества", задающаяся перестановкой $p(1)p(2)\ldots p(\tau)$:

$$q_{j,p(1)} \geq q_{j,p(2)} \geq \cdots \geq q_{j,p(\tau)}.$$

- Отбор осуществляется выбором $K$ кандидатов $\mathbf{I}_{p(1)}, \ldots, \mathbf{I}_{p(K)}$ с наибольшим значением качества.

## Методы оценки качества

- Оценка качества *(L. Best-Rowden)* [*Deep-QA-SORT*]
  - обучение с учителем, построение регрессии признакового представления объектов на асессорскую оценку качества, модель SVM;
  - признаковое представление извлекается сверточной нейронной сетью, подобной сети ре-идентификации;
- Оценка качества на основе уверенности (confidence) детектора

## Показатели качества

- Precision – точность;
- Recall – полнота;
- IDS – суммарное число некорректных переключений при продлении траекторий;
- Hz – частота работы (в кадрах в секунду).

## Базы данных

1. MOT20-01 – данные камеры наблюдения, высокий ракурс;
2. MOT20-02 – данные камеры наблюдения, средний ракурс;

| Выборка | FPS | Плотность на кадр | Длина | Траектории |
|---|---|---|---|---|
| MOT20-01 | 25 | 42.1 | 429 | 90 |
| MOT20-02 | 25 | 72.7 | 2782 | 296 |

Описание выборок

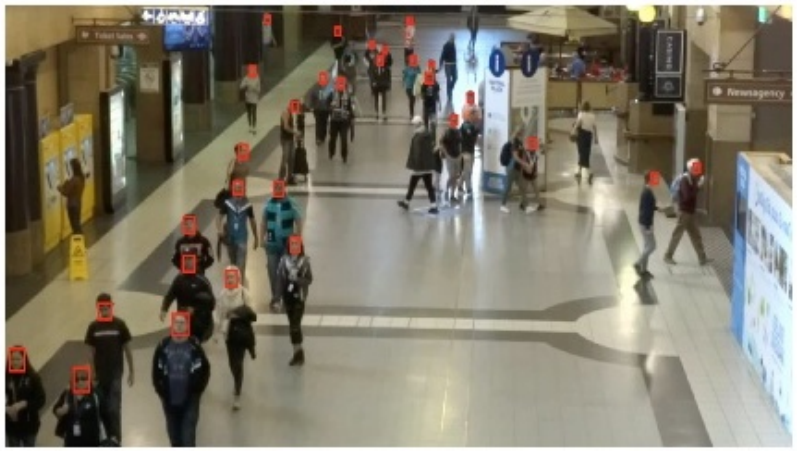
## Локализация лиц

Для выделения областей, содержащих объекты, используются детекторы разной сложности: SSD и RetinaNet.

| Архитектура | Число обучаемых параметров |
|---|---|
| SSD | 4 M |
| RetinaNet | 45 M |

MOT20-01

MOT20-02



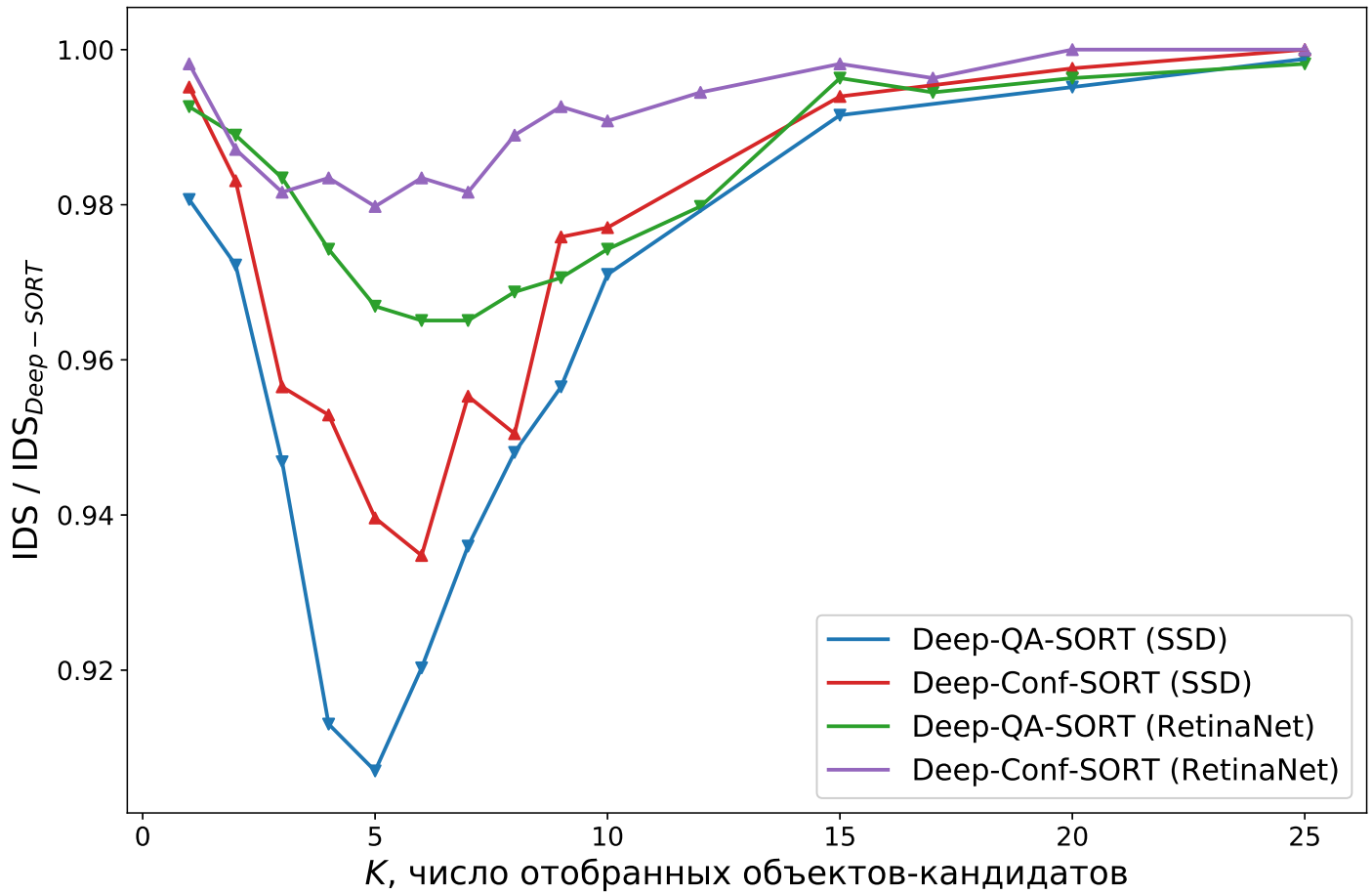
Примеры локализаций лиц на кадрах из выборок MOT20-01, MOT20-02

# Результаты эксперимента ($K = 5$)

| Детектор | Метод слежения | Precision | Recall | IDS | Hz |
|---|---|---|---|---|---|
| выборка MOT20-02 | | | | | |
| SSD | Deep-SORT | 0.851 | **0.910** | 828 | 22.6 |
| | Deep-Conf-SORT | **0.860** | 0.904 | 778 | **31.8** |
| | Deep-QA-SORT | 0.855 | 0.907 | **751** | 31.8* |
| RetinaNet | Deep-SORT | 0.887 | **0.945** | 543 | 22.4 |
| | Deep-Conf-SORT | 0.891 | 0.941 | 533 | **31.4** |
| | Deep-QA-SORT | **0.896** | 0.939 | **526** | 31.4* |
| выборка MOT20-01 | | | | | |
| SSD | Deep-SORT | 0.847 | **0.895** | 203 | 39.2 |
| | Deep-Conf-SORT | **0.850** | 0.890 | 202 | **44.7** |
| | Deep-QA-SORT | 0.848 | 0.892 | **195** | 44.7* |
| RetinaNet | Deep-SORT | 0.871 | **0.922** | 168 | 40.9 |
| | Deep-Conf-SORT | **0.875** | 0.919 | 166 | **46.0** |
| | Deep-QA-SORT | 0.873 | 0.921 | **162** | 46.0* |

* Время работы алгоритма оценки качества не учтено

# Результаты эксперимента




Зависимость отношения IDS метода к IDS базового метода от числа отобранных кандидатов $K$

# Результаты, выносимые на защиту

- Предложены методы слежения за множеством объектов в видеопотоке, использующие процедуру отбора объектов на основе оценки качества для ре-идентификации.
- Введен метод оценки качества на основе уверенности детектора, не требующий дополнительных вычислений.
- Показана эффективность предложенных методов в терминах числа некорректных переключений при продлении траекторий при увеличении скорости работы алгоритма по сравнению со стандартным подходом.