Использование глубоких сверточных нейронных сетей для извлечения визуальных признаков в задаче отслеживания объектов на видеопоследовательности

А.Е. Мещеряков[1], С.Б. Попов[1]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. В статье исследуется задача сопоставления визуальных признаков при отслеживании объектов на видеопоследовательности. Авторами проведен сравнительный анализ существующих методов извлечения визуальных признаков объектов и оценки сходства признаков (similarity estimation) для повторной идентификации объектов. Разработан программный комплекс, в котором реализовано несколько алгоритмов извлечения признаков и оценки их сходства. Авторами проведена экспериментальная оценка скорости и точности работы алгоритмов с помощью наборов данных MOT-16 и PETS 2007. Показано, что наиболее точные оценки сходства объектов достигаются при расчете модифицированного значения нормализованной взаимнокорреляционной функции между признаками, извеченными из слоя субдискретизации нейронной сети.

1. Введение

Повсеместное распространение систем видеонаблюдения стало толчком к разработке и развитию алгоритмов отслеживания, или трекинга, объектов. Более 70% всех исследований, посвященных отслеживанию нескольких объектов, так или иначе связаны с отслеживанием людей [1].

Существует несколько подходов к классификации алгоритмов отслеживания нескольких объектов. Наиболее распространенной является классификация на алгоритмы, использующие детектор объектов (detection-based tracking, DBT) и не использующие его (detection-free tracking, DFT) [2]. DBT-алгоритмы являются более распространенными, поскольку они позволяют автоматически обнаруживать новые объекты в кадре и определять объекты, покинувшие кадр [2]. Однако при этом производительность и точность всего алгоритма отслеживания зависят от производительности и точности детектора, поэтому к алгоритму детектирования выдвигаются особые требования по времени обработки кадров.

Кроме того, одной из важнейших частей любого алгоритма отслеживания объектов является этап повторной идентификации, или сопоставления, объектов. На данном этапе для каждого кадра видеопоследовательности выполняется извлечение визуальных признаков обнаруженных объектов и их сопоставление с признаками уже идентифицированных объектов на предыдущем кадре для идентификации объектов на текущем кадре. При этом в процессе сопоставления объектов с идентификаторами может возникнуть ошибка, когда объекту присваивается идентификатор другого объекта. Такая ситуация называется переключением идентификатора

(ID switch) и является одной из самых серьезных ошибок при отслеживании объектов [3]. Поэтому к методам извлечения признаков обнаруженных объектов и оценки их сходства предъявляются дополнительные требования, связанные с устойчивостью этих методов к ошибкам переключения идентификаторов.

В данной статье приводится сравнение различных методов извлечения визуальных признаков объектов, обнаруженных детектором, а также оценка различных метрик расстояния между признаками. В рамках данной работы не исследуются возможные подходы к детектированию объектов, и делается предположение, что для этапа повторной идентификации каждого кадра известны объекты, присутствующие на этом кадре, и определены координаты их обрамляющий прямоугольников (bounding box).

2. Постановка задачи

Для постановки задачи извлечения и сопоставления визуальных признаков сначала требуется дать формулировку задачи отслеживания объектов. Пусть исходная видеопоследовательность дана в виде последовательности кадров $T = \{t_1, t_2, \dots, t_n\}$. $M = \{m_1, m_2, \dots, m_k\}$ – множество всех отслеживаемых объектов на видеопоследовательности, $S_t = \{s_t^{m_1}, s_t^{m_2}, \dots, s_t^{m_k}\}$ – множество состояний всех объектов на кадре t (при этом на каждом отдельно взятом кадре могут быть представлены не все объекты, и, следовательно, некоторые элементы S_t будут пустыми) [1]. Таким образом, $S_{i_s:i_e}^{m_i} = \left\{s_{i_s}^{m_i}, \dots, s_{i_e}^{m_i}\right\}, 1 \le i \le k, 1 \le i_s \le i_e \le n$ – упорядоченное множество состояний объекта m_i, где i_s и i_e – первый и последний кадр видеопоследовательности, соответственно, на которых присутствует объект m_i. $S_{1:n}$ – упорядоченное множество всех состояний всех объектов из M от кадра 1 до кадра n. Поскольку в данной статье рассматривается DBT-трекинг, вводится множество O_t детектирований (или наблюдений) всех объектов на кадре t, и $O_{1:n}$ – упорядоченное множество наблюдений всех объектов из M от кадра 1 до кадра n. Исходя из этого, можно сформулировать задачу отслеживания объектов как задачу оценки апостериорного максимума условного распределения состояний $S_{1:n}$ при известных $O_{1:n}$ [1]:

$$\hat{S}_{1:n} = \arg\max_{S_{1:n}} P(S_{1:n} | O_{1:n}) \quad \#(1)$$

Извлечением признаков из наблюдения o называется отображение этого наблюдения на p-мерное пространство признаков X^p:

$$f: O_{1:n} \to X^p \quad \#(2)$$

Здесь f – функция извлечения признаков. Основной особенностью данной функции является то, что отображения наблюдений одного и того же объекта, полученные на разных кадрах, должны находиться как можно ближе друг к другу в X^p :

$$\forall\, i \in [1..k]\ \forall\, n_1, n_2 \in [1..n]\ \ d\left(f\left(o_{n_1}^{m_i}\right), f\left(o_{n_2}^{m_i}\right)\right) \to 0\,, \#(3)$$

где $d(x_1, x_2)$ – некая метрика расстояния между двумя отображениями.

2.1. Существующие методы извлечения признаков

В качестве функции извлечения признаков могут быть использованы разные методы извлечения информации из изображения. В работе [1] такие методы делятся на две группы: методы локальных признаков и методы региональных признаков. Типичным методом локальных признаков является извлечение признаков Канаде-Лукаса-Томаси (KLT features) [4]. Также распространено использование оптического потока для кодирования информации о движении объектов и выстраивания объектов, обнаруженных детектором, в частичные треки, или треклеты [5, 6, 7]. Одним из больших преимуществ методов локальных признаков является возможность их применения для определения паттернов движения объектов в переполненном (overcrowded) кадре [8].

Методы региональных признаков, в отличие от методов локальных признаков, обрабатывают большую часть изображения кадра. Методы региональных признаков можно разделить на методы нулевого, первого и высших порядков [1]. К методам нулевого порядка

можно отнести метод построения гистограммы цветов [7, 9]. Здесь нулевой порядок метода означает, что метод оперирует «сырыми» значениями пикселей, не сравнивая их с другими. Недостатком такого метода является то, то он игнорирует пространственные признаки объекта, к примеру, взаимное расположение его частей и т.д. К методам первого порядка относятся методы, связанные с использованием градиентов каких-либо значений пикселей, к примеру, гистограмма ориентированных градиентов (HOG) [7, 9, 10]. К методам высших порядков относятся методы, использующие различия второго порядка и выше, к примеру, использование матрицы ковариации [11]. Методы высших порядков являются более устойчивыми, однако, недостатком применения таких методов является возрастание времени вычислений [1].

В последнее время глубокие сверточные нейронные сети демонстрируют крайнюю эффективность в решении задач компьютерного зрения, в том числе классификации изображений и детектирования объектов [12]. Также, базируясь на архитектурах и принципах современных нейронных сетей, в настоящее время стремительно развивается область переноса обучения (transfer learning), в которой особую роль играет грамотная интерпретация признаков, извлеченных нейронной сетью [13]. Также обучение с переносом знания позволяет обеспечить нейронные сети высокой дискриминативной силой (в том смысле, что нейронная сеть способна различать как объекты, принадлежащие разным классам, так и извлекать разные признаки для разных объектов, принадлежащих одному классу) [14]. Так, в работе [15] авторы используют механизм переноса знаний для детектирования объектов извлечения их признаков из выходов сверточных слоев нейронной сети, при этом такой подход на текущий момент является одним из наиболее точных. Поэтому в настоящей работе будет исследован метод извлечения признаков из выходов сверточных слоев нейронной сети.

2.2. Существующие методы оценки сходства признаков

В качестве метрик расстояния между отображениями признаков чаще всего используется расстояние Минковского с параметрами, равными 1 и 2 (манхэттенское и евклидово расстояние соответственно) [16]. Кроме того, для оценки близости признаков используется значение нормализованной взаимнокорреляционной функции [8, 17] и коэффициент Бхаттачарии [18]. В данной работе будет проведено сравнение двух последних метрик.

3. Предлагаемый алгоритм

На первом шаге обработки каждого кадра видеопоследовательности T необходимо определить находящиеся на нем объекты, после чего извлечь из них признаки. Для определения положения объектов на кадре возможно использовать любой алгоритм детектирования объектов, в том числе основанный на нейронных сетях. Важно, чтобы границы каждого объекта были четко определены, к примеру, с применением метода обрамляющего прямоугольника, как это сделано в работе [19].

Для извлечения признаков из обнаруженных и разграниченных объектов предлагается использовать глубокую сверточную нейронную сеть (СНС) архитектуры SqueezeNet [20]. Схема расположения слоев данной нейронной сети приведена на рисунке 1. Преимуществами данной архитектуры, по сравнению с нейронными сетями семейства VGG, являются гораздо меньшее количество параметров (и, как следствие, время обработки изображения) и сравнимая точность классификации [20]. Поскольку основным требованием к нейронной сети, предназначенной для извлечения признаков, является ее высокая дискриминативная сила, применялся алгоритм обучения нейронной сети с переносом знания. Предварительное обучение нейронной сети проведено на наборах данных ImageNet и OpenImages для классификации изображений, однако допустимо проводить предварительное обучение на других наборах данных, в которых множество целевых классов включает в себя классы отслеживаемых объектов.

После предварительного обучения нейронная сеть может быть использована для извлечения признаков. Для этого на вход нейронной сети подается вырезанное из кадра изображение отслеживаемого объекта, отмасштабированное до размера 224×224 пиксела. Извлеченные признаки формируются из выходов слоя *global avgpool* и имеют размерность 1000. Таким

образом, нейронная сеть, согласно формуле (2), представляет собой отображение изображений объектов в 1000-мерное пространство признаков X^{1000}.

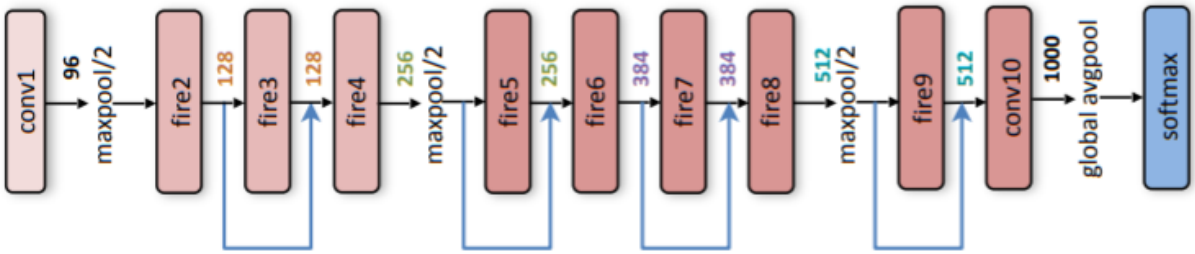

Рисунок 1. Структура нейронной сети SqueezeNet [20].

На следующем этапе необходимо вычислить метрику дистанции между извлеченными на текущем кадре признаками объектов и уже известными признаками идентифицированных объектов для идентификации объектов на текущем кадре. Предлагается использовать следующии функции в качестве метрик:

1. Модифицированное значение нормализованной взаимнокорреляционной функции:

$$d_{\text{ncc}} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_{x_t}\sigma_{x_{t-1}}}\left(x_{t_i} - \mu_{x_t}\right)\left(x_{t_{i-1}} - \mu_{x_{t-1}}\right), \#(4)$$

где n – размерность пространства признаков (в данном случае 1000), x_t – извлеченные признаки какого-либо объекта на текущем кадре, x_{t-1} – извлеченные признаки какого-либо идентифицированного объекта на предыдущем кадре, μ_{x_t} и σ_{x_t} – среднее значение и стандартное отклонение, соответственно, признаков x_t, $\mu_{x_{t-1}}$ и $\sigma_{x_{t-1}}$ – среднее значение и стандартное отклонение, соответственно, признаков x_{t-1}. d_{ncc} может принимать значения из интервала [0;1], где 0 означает что признаки совпадают, 1 – не совпадают.

2. Модифицированное значение коэффициента Бхаттачарии:

$$d_{\text{B}} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(x_{t_i}x_{t_{i-1}}\right)^{1/2}, \#(5)$$

где n – размерность пространства признаков (в данном случае 1000), x_t – извлеченные признаки какого-либо объекта на текущем кадре, x_{t-1} – извлеченные признаки какого-либо идентифицированного объекта на предыдущем кадре. d_{B} также принимает значения из интервала [0;1], где 0 означает что признаки совпадают, 1 – не совпадают.

Общая схема предлагаемого программного комплекса представлена на рисунке 2. Исходная видеопоследовательность обрабатывается покадрово. На первом этапе происходит предварительная обработка кадра (нормализация, удаление шума). Затем на кадре с помощью детектора объектов выполняется поиск отслеживаемых объектов. После того, как все отслеживаемые объекты найдены, из кадра вырезаются обрамляющие прямоугольники объектов, которые затем масштабируются до размера 224×224 пиксела. Отмасштабированные изображения объектов передаются в нейронную сеть, на выходе из которой извлекаются признаки объектов. После этого для каждого из признаков на текущем кадре вычисляется расстояние до каждого из признаков уже идентифицированных объектов предыдущего кадра. Затем система, основываясь на величине вычисленных расстояний, и с учетом формулы (3), решает задачу о назначении каждому из признаков текущего кадра идентификаторов объектов, обнаруженных на предыдущем кадре. После идентификации объектов текущего кадра их признаки и информация об идентификаторах вносится в хранилище признаков, и система приступает к обработке следующего кадра.

Для сопоставления признаков используется венгерский алгоритм, решающий задачу о назначениях признакам текущего кадра признаков предыдущего кадра таким образом, что сумма расстояний между признаками должна быть минимальной [21].

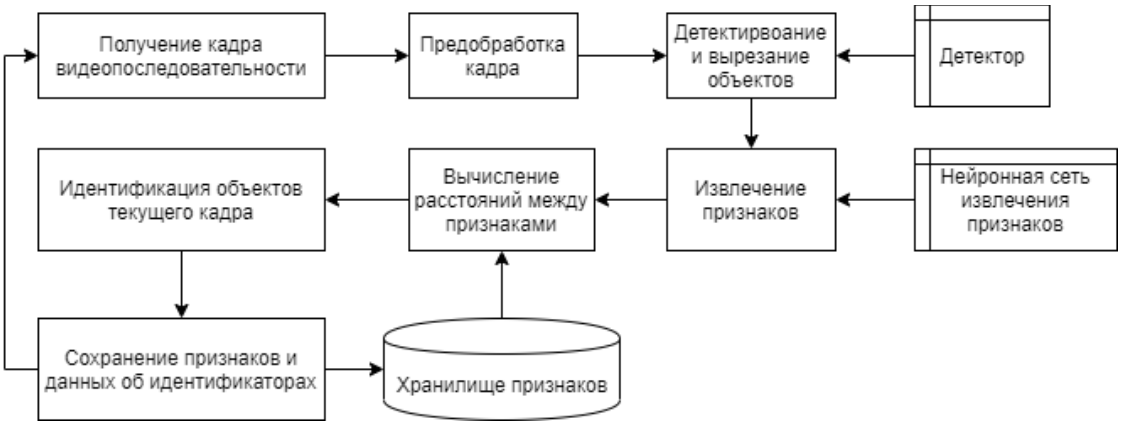

Рисунок 2. Схема программного комплекса.

4. Результаты экспериментов

Описанный программный комплекс был реализован на языке программирования Python 3.7.0 с применением библиотеки OpenCV и фреймворка глубокого обучения Keras. В качестве хранилища признаков использована СУБД PostgreSQL 10.5. Все описанные ниже эксперименты проводились на ПЭВМ следующей конфигурации: CPU Intel Core i5-8350, 16Гб DDR4 RAM, GPU Nvidia GeForce GTX 1070.

Для проведения экспериментов использовались размеченные наборы видеопоследовательностей MOT-16 [22] и PETS 2007 [23]. Характеристики этих видеопоследовательностей приведены в таблице 1. Для каждого кадра тестовых видеопоследовательностей были определены положения отслеживаемых объектов и их идентификаторов.

Таблица 1. Характеристики наборов видеопоследовательностей MOT-16 и PETS 2007.

	MOT-16	PETS 2007
Разрешение кадров	1920x1080	768x576
Частота кадров (FPS)	30	25
Минимальная высота отслеживаемого объекта, пикселей	19	19
Максимальная высота отслеживаемого объекта, пикселей	556	230

В таблице 2 приведено среднее время извлечения признаков с помощью нейронной сети SqueezeNet, описанной в разделе 3, на ПЭВМ с использованием и без использования GPU. Для сравнения, было измерено среднее время извлечения признаков из аналогичных изображений объектов методом построения гистограммы цветов для каждого из трех каналов изображения с количеством бинов, равным 256, по методике, описанной в работе [9]. Наилучшие результаты выделены полужирным шрифтом.

Таблица 2. Среднее время извлечения признаков (сек).

	SqueezeNet	Гистограмма цветов
С использованием GPU (1 изображение)	0.021	**0.012**
Без использования GPU (1 изображение)	0.245	**0.026**
С использованием GPU (16 изображений)	**0.023**	0.203
Без использования GPU (16 изображений)	**0.309**	0.487

Из таблицы 2 видно, что при извлечении признаков из изображения одного объекта и использовании только CPU нейронная сеть в несколько раз медленнее построения гистограммы

цветов, однако при использовании GPU время обработки этого изображения сравнимо со временем построения гистограммы цветов. Однако при проведении пакетной обработки изображений (batch processing) время, затрачиваемое нейронной сетью, растет незначительно, в то время как время на построение гистограммы цветов растет линейно. Поскольку в задачах отслеживания объектов обычно в кадре находится сразу несколько объектов, использование нейронной сети оправдано.

В таблице 3 представлены результаты экспериментальной проверки возможности использования модифицированного значения взаимнокорреляционной функции и модифицированного коэффициента Бхаттачарии для оценки расстояния между признаками и их последующей повторной идентификации. В качестве отслеживаемых объектов рассматривались только пешеходы. Извлечение признаков проводилось с применением нейронной сети SqueezeNet. Результаты этих методов сравниваются с точностью извлечения и сопоставления признаков методами гистограммы цветов и гистограммы направленных градиентов. Поскольку для каждой тестовой видеопоследовательности заранее известно множество M объектов, присутствующих на ней, и, следовательно, известно общее количество объектов, задача сопоставления признаков может быть рассмотрена как задача классификации признаков на k непересекающихся классов, где каждому классу соответствует один и только один объект из множества M. В качестве целевой метрики качества повторной идентификации использовалась точность (accuracy) классификации.

Одной из основных метрик качества алгоритма отслеживания объектов является количество ошибок смены идентификаторов (ID Switch), когда отслеживаемый и ранее идентифицированный объект на некоторое время перекрывается препятствием (к примеру, другим отслеживаемым объектом), и после появления в поле видимости камеры ему присваивается новый идентификатор [24]. Понятно, что алгоритмы извлечения и сопоставления визуальных признаков должны стремиться минимизировать количество ошибок смены идентификаторов. В таблице 3 также приведены результаты измерения количества ошибок смены идентификаторов на наборе данных MOT-16 при извлечении признаков с помощью сверточной нейронной сети (с использованием в качестве метрики сходства модифицированного значения нормализованной взаимокорреляционной функции и модифицированного значения коэффициента Бхаттачарии), с помощью гистограммы цветов и гистограммы направленных градиентов.

Таблица 3. Средняя точность (accuracy) сопоставления признаков.

	MOT-16	PETS 2007	ID Switch
СНС + модифицированное значение нормализованной взаимокорреляционной функции (4)	**0.695**	**0.752**	**1503**
СНС + модифицированное значение коэффициента Бхаттачарии (5)	0.611	0.726	1810
Гистограмма цветов	0.341	0.448	2418
Гистограмма направленных градиентов	0.357	0.487	3012

5. Заключение

В статье исследована возможность использования сверточных нейронных сетей для извлечения признаков отслеживаемых объектов при решении задачи отслеживания объектов, а также приведены две метрики дистанции между извлеченными признаками в пространстве признаков.

Использование модифицированного значения нормализованной взаимнокорреляционной функции показало наибольшую эффективность при решении задачи повторной идентификации объектов по извлеченным из них признакам. При этом стоит отметить, что использование сверточной нейронной сети для извлечения признаков при использовании GPU и единовременной обработке сразу нескольких объектов дает существенный выигрыш во времени обработки изображений объектов.

6. Литература

[1] Luo, W. Multiple object tracking: A literature review // arXiv preprint arXiv:1409.7618, 2014.
[2] Yang, B. Online learned discriminative partbased appearance models for multi-human tracking / B. Yang, R. Nevatia // Proc. Eur. Conf. Comput. Vis, 2012. – P. 484-498.
[3] Yamaguchi, K. Who are you with and where are you going? / K. Yamaguchi, A.C. Berg, L.E. Ortiz, T.L. Berg // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. – 2011. – P. 1345-1352.
[4] Lucas, B.D. An Iterative Image Registration Technique with an Application to Stereo Vision / B.D. Lucas, T. Kanade // International Joint Conference on Artificial Intelligence. – 1981. – P. 674-679.
[5] Walk, S. New features and insights for pedestrian detection / S. Walk, N. Majer, K. Schindler, B. Schiele // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. – 2010. – P. 1030-1037.
[6] Rodriguez, M. Tracking in unstructured crowded scenes / M. Rodriguez, S. Ali, T. Kanade // Proc. IEEE Int. Conf. Comput. Vis, 2009. – P. 1389-1396.
[7] Izadinia, H. (MP)2T: Multiple people multiple parts tracker / H. Izadinia, I. Saleemi, W. Li, M. Shah // Proc. Eur. Conf. Comput. Vis, 2012. – P. 100-114.
[8] Ali, S. Floor fields for tracking in high density crowd scenes / S. Ali, M. Shah // Proc. Eur. Conf. Comput. Vis, 2008. – P. 1-14.
[10] Sugimura, D. Using individuality to track individuals: Clustering individual trajectories in crowds using local appearance and frequency trait / D. Sugimura, K.M. Kitani, T. Okabe, Y. Sato, A. Sugimoto // Proc. IEEE Int. Conf. Comput. Vis, 2009. – P. 1467-1474.
[11] Porikli, F. Covariance tracking using model update based on lie algebra / F. Porikli, O. Tuzel, P. Meer // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, 2006. – P. 728-735.
[12] Image Classification on ImageNet [Электронный ресурс]. – Режим доступа: https://paperswithcode.com/sota/image-classification-on-imagenet (03.11.2019).
[13] Tan, C. A survey on deep transfer learning / C. Tan // International Conference on Artificial Neural Networks, 2018. – P. 270-279.
[14] Zhao W. A Multisubregion-Based Probabilistic Approach Toward Pose-Invariant Face Recognition / W. Zhao, R. Chellappa – Cambridge: Academic Press, 2006. – 738 p.
[15] Wang Z. Towards Real-Time Multi-Object Tracking // arXiv preprint arXiv:1909.12605, 2019.
[16] Yu, J. Distance Learning for Similarity Estimation / J. Yu, J. Amores, N. Sebe, P. Radeva, Q. Tian // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2008. – Vol. 30(3). – P. 1-10.
[17] Wu, Z. Coupling detection and data association for multiple object tracking / Z. Wu, A. Thangali, S. Sclaroff, M. Betke // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. – 2012. – P. 1948-1955.
[18] Xing, J. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses / J. Xing, H. Ai, S. Lao // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2009. – P. 1200-1207.
[19] Lee, S. Universal Bounding Box Regression and Its Applications / S. Lee, S. Kwak, M. Cho // Asian Conference on Computer Vision, 2018. – P. 373-387.
[20] Iandola, F.N. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size // arXiv preprint arXiv:1602.07360, 2016.
[21] Perera, A.A. Multi-object tracking through simultaneous long occlusions and split-merge conditions / A.A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, W. Hu // Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2006. – P. 666-673.
[22] MOT Challenge [Electronic resource]. – Access mode: https://motchallenge.net/ (03.11.2019).
[23] PETS 2007 [Electronic resource]. – Access mode: http://www.cvg.reading.ac.uk/PETS2007/index.html (03.11.2019).
[24] Milan, A. MOT16: A benchmark for multi-object tracking // arXiv preprint arXiv:1603.00831, 2016.

Using of deep convolutional neural networks for visual features extraction in multiple objects tracking task

A.E. Meshcheriakov[1], S.B. Popov[1]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

Abstract. The article explores the task of comparing visual features when tracing objects on a video sequence. The authors conducted a comparative analysis of existing methods for extracting visual signs of objects and assessing the similarity of signs (similarity estimation) for re-identification of objects. A software package has been developed that implements several algorithms for extracting attributes and assessing their similarity. The authors carried out an experimental assessment of the speed and accuracy of the algorithms using the MOT-16 and PETS 2007 datasets. It is shown that the most accurate estimates of the similarity of objects are achieved by calculating the modified value of the normalized cross-correlation function between features derived from the neural network average pooling layer.