Данный текст является русскоязычной версией опубликованной на английском языке статьи и представлен в авторской редакции только на данном сайте!

UDC 004.89
Doi: 10.31772/2587-6066-2020-21-2-150-154

For citation: Buryachenko V. V., Pahirka A. I. Object tracking with deep learning. *Siberian Journal of Science and Technology*. 2020, Vol. 21, No. 2, P. 150–154. Doi: 10.31772/2587-6066-2020-21-2-150-154

Для цитирования: Буряченко В. В., Пахирка А. И. Методы слежения за объектами с применением глубокого обучения // Сибирский журнал науки и технологий. 2020. Т. 21, № 2. С. 150–154. Doi: 10.31772/2587-6066-2020-21-2-150-154

МЕТОДЫ СЛЕЖЕНИЯ ЗА ОБЪЕКТАМИ С ПРИМЕНЕНИЕМ ГЛУБОКОГО ОБУЧЕНИЯ

В. В. Буряченко*, А. И. Пахирка

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
*E-mail:buryachenko@sibsau.ru

Слежение за объектами является ключевой задачей видеоаналитики и компьютерного зрения, которая имеет множество применений в различных областях. Большинство систем слежения включают в себя два этапа: обнаружение объектов и отслеживание изменения положения объектов. На первом этапе осуществляется обнаружение объектов интереса в каждом кадре видеопоследовательности, а на втором выполняется оценка соответствия обнаруженных объектов в соседних кадрах. Тем не менее в сложных условиях видеонаблюдения данная задача имеет ряд особенностей, связанных с изменением освещенности кадра, изменением формы объектов, например при ходьбе человека, а также усложняется в случае движения камеры. Целью работы является разработка метода слежения за объектами на основе нейронных сетей глубокого обучения, который позволяет осуществлять отслеживание нескольких объектов в кадре, в том числе и в сложных условиях видеонаблюдения. В работе выполнен обзор современных методов решения задач слежения за объектами, среди которых наиболее перспективным подходом является использование сетей глубокого обучения. Основным используемым подходом в данной статье являются нейронные сети для обнаружения регионов (R-CNN), которые показали себя эффективным методом для решения задач обнаружения и распознавания объектов на изображениях. Предложенный алгоритм использует ансамбль, содержащий две глубокие нейронные сети для обнаружения объектов и уточнения результатов классификации и выделения границ объекта. В статье выполнена оценка эффективности разработанной системы с использованием классической метрики MOT в области слежения за объектами на известных базах данных, доступных в открытых источниках. Проведено сравнение эффективности предложенной системы с другими известными работами.

Ключевые слова: интеллектуальные системы, глубокое обучение, оценка движения, сверточная сеть для классификации регионов.

OBJECT TRACKING WITH DEEP LEARNING

V. V. Buryachenko*, A. I. Pahirka

Reshetnev Siberian State University of Science and Technology
31, Krasnoyarskii rabochii prospekt, Krasnoyarsk, 660037, Russian Federation
*E-mail: buryachenko@sibsau.ru

Tracking objects is a key task of video analytics and computer vision, which has many applications in various fields. Most tracking systems include two steps: detecting objects and tracking changes in the position of objects. At the first stage, objects of interest are detected in each frame of the video sequence, and at the second, the correspondence of the detected objects in neighboring frames is assessed. Nevertheless, in difficult conditions of video surveillance, this task has a number of difficulties associated with changing the illumination of the frame, changing the shape of objects, for example, when a person is walking, and the task is also complicated in the case of camera movement. The aim of the work is to develop a method for tracking objects on the basis of deep learning, which allows you to track several objects in the frame, including in the difficult conditions of video surveillance. The paper provides an overview of modern methods for solving objects tracking tasks, among which the most promising approach of the using of deep learning neural networks. The main approach used in this paper is neural networks for detecting regions (R-CNN), which have proven to be an effective method for solving problems of detection and recognition of objects in images. The proposed algorithm uses an ensemble containing two deep neural networks to detect objects and to refine the results of classification and highlight the boundaries of the object. The article evaluates the effectiveness of the developed system using the classic in the field MOT metric for objects tracking on known databases available in open sources. The effectiveness of the proposed system is compared with other well-known works.

Keywords: intelligent systems, deep learning, motion estimation, convolutional network for the regions classification (R-CNN).

Введение. В последние годы методы оценки движения и отслеживания объектов достигли впечатляющих результатов. Практическое применение в целях повышения безопасности имеют работы по слежению за передвижением людей в общественных местах, методы распознавания лиц, обнаружение подозрительных объектов и отклоняющегося от стандартов поведения людей. Методы стабилизации видеопоследовательностей также широко используются в задачах видеоаналитики и для повышения удобства операторов при работе систем видеонаблюдения. Стабилизация позволяет устранить непреднамеренное дрожание видео, сохранив при этом направленное движение и панорамирование камеры.

Большинство из рассмотренных задач основано на применении методов высокоуровневого анализа видеопоследовательностей, оценки движения и слежения за объектами, для чего требуется выполнить ресурсоемкие задачи по обнаружению и отслеживанию особенных точек в кадрах видеопоследовательности. Предлагается использовать современные технологии глубокого обучения для повышения качества и уменьшения вычислительных и временных затрат на выполнение оценки движения и стабилизации видеопоследовательностей.

Обзор литературы. Глубокие нейронные сети показали себя одной из лучших технологий для решения множества задач, связанных с обработкой графической информации. Наибольшие успехи продемонстрировали сверточные нейронные сети, содержащие от 10 до 300 слоев, при решении задач распознавания изображений [1], семантическом анализе текстов [2] и обучении с подкреплением [3]. Недавние исследования показали эффективность искусственных нейронных сетей сложных структур в решении задач анализа

движения [4; 5] и оценки оптического потока [6; 7], а также возможность применения таких технологий для стабилизации видеопоследовательностей [8]. Одним из известных подходов, использующихся для обнаружения и отслеживания объектов, является оценка заметности в кадре [9; 10].

В данной работе рассмотрены методы слежения за движущимися объектами с применением подходов, основанных на глубоком обучении. Для повышения качества алгоритма применяется сверточная сеть для классификации регионов (R-CNN) и методы стабилизации полученного видеоматериала. Визуальное слежение за объектами является классической задачей компьютерного зрения, в которой положение цели определяется в каждом кадре. Данная область исследования остается востребованной в связи с большим количеством практических задач, основанных на слежении за различными объектами. Алгоритмы, применяющиеся в этой сфере, также улучшаются и позволяют решать все более сложные проблемы, такие как окклюзия, изменение положения и формы объектов, внешнего вида людей, изменение освещения и наличие сложного фона с различной текстурой. В связи с этим, в литературе предлагается множество алгоритмов и подходов, направленных на решение различных проблем слежения и улучшение общей производительности отслеживания объектов.

Слежение за объектами. Типичная система отслеживания состоит из двух основных моделей: движения и внешнего вида. Модель движения используется для прогнозирования местоположения цели в следующем кадре, как при использовании фильтра Калмана или фильтра частиц для моделирования движения цели. Модель движения также может быть как простой, например, включающей линейное движение, на основе которого строятся несколько более сложных траекторий, так и более сложной, выполняющей отслеживание объектов с учетом изменения направления и скорости движения. Для ускорения процесса оценки движения предлагается допущение о величине движения в рамках окна поиска вокруг предыдущего местоположения объекта. С другой стороны, модель внешнего вида используется для описания цели и проверки предсказанного местоположения цели в каждом кадре. Модели внешнего вида могут использовать генеративные и дискриминационные методы. В генеративных методах отслеживание выполняется путем поиска наиболее похожего на объект региона. В дискриминационных методах используется классификатор, позволяющий отличать объект от фона. В общем случае модель внешнего вида может обновляться в процессе работы системы для того, чтобы учитывать необходимые изменения объектов. Это позволяет, например, продолжать слежение за человеком при повороте или наклоне тела.

Традиционно в алгоритмах отслеживания движения использовались рассчитанные вручную функции, основанные на интенсивности пикселей, цвете и гистограмме ориентированных градиентов (Histogram oriented gradient, HOG), для представления цели в моделях генеративного или дискриминационного описания. Хотя они достигают удовлетворительных характеристик в определенных условиях, они не устойчивы к серьезным изменениям внешнего вида объектов. Глубокое обучение с использованием Convolution Neural Networks (CNN) недавно позволило значительно повысить производительность различных приложений компьютерного зрения.

Этот подход также затронул визуальное отслеживание объектов и частично позволил преодолеть сложности и получить лучшую производительность по сравнению с методами, применявшимися ранее. В системах отслеживания на основе CNN модель внешнего вида объекта строится на основе обучения сверточной сети, а классификатор используется для маркирования пути на изображении как принадлежащего объекту или фону. Системы на основе CNN достигли современного уровня эффективности даже с использованием простых моделей движения в автономном режиме без переобучения. Однако такие системы обычно

испытывают высокие вычислительные нагрузки из-за большого количества возможных траекторий объектов на этапах обучения нейронной сети и отслеживания объектов.

Перспективным направлением в области слежения за объектами являются задачи, связанные с анализом большого числа людей в кадре (Multi-object tracking, MOT). Эта задача имеет два этапа: обнаружение объектов и ассоциация их в разных кадрах. На первом этапе нужные объекты обнаруживаются в каждом кадре видеопотока, причем в зависимости от используемого детектора эти объекты могут быть различны. Качество обнаружения напрямую влияет на производительность системы слежения. Второй этап включает в себя поиск соответствия обнаруженных объектов в текущем кадре с предыдущим для оценки их траекторий движения. Высокая точность системы обнаружения объектов приводит к меньшему количеству пропущенных объектов и к получению более стабильных траекторий. Но при этом чрезмерное переобучение системы обнаружения может снизить качество системы при изменении параметров самих объектов. Применение дополнительных подходов при решении задачи ассоциации объектов в разных кадрах может повысить качество слежения за сложными объектами с изменяющимися параметрами. Точность обнаружения объектов может быть увеличена за счет использования сверточной нейронной сети на основе глубокого обучения. Объединение объектов осуществляется с учетом внешнего вида и улучшенных функций движения.

Система для слежения за объектами на основе R-CNN. Одним из направлений исследования является разработка алгоритма, позволяющего повысить точность отслеживания объектов за счёт применения сверточной сети для классификации регионов (R-CNN) и увеличить скорость работы системы до показателей близких к реальному времени. Сверточная сеть позволяет классифицировать область на принадлежность объекту или фону, и при этом карта признаков, полученная при работе сети, также используется для выполнения приближенной локализации объектов и позволяет сократить время на оценку соответствий между ними.

Практические эксперименты проводились с использованием нейронной сети для классификации регионов, структура которой приведена в табл. 1. Основной задачей является отслеживание нескольких типов объектов в кадре: людей и машин. Предложенный метод включает в себя компоненты обнаружения людей, прогнозирования положения объектов в следующих кадрах, ассоциации обнаруженных объектов и управления циклом выделенных объектов. Наиболее широко известный алгоритм обнаружения объектов – YOLO [24], который достаточно быстр для обнаружения нескольких объектов в реальном времени, но имеет недостаточную точность, что приводит к фрагментации траекторий и сложностям с идентификацией обнаруженных объектов.

Основная схема разработанной системы представлена на рис. 1. С развитием алгоритмов, основанных на глубоком обучении, обнаружение объектов в сложных условиях стало на много проще. Ключевым компонентом алгоритма является сверточная сеть для обнаружения регионов (R-CNN).

Таблица 1

Параметры сверточной нейронной сети

Название слоя	Размер фильтра	Шаг	Количество выходов слоя
Conv 1	3 × 3	1	32×128×6432×128×64
Conv 2	3 × 3	1	32×128×6432×128×64
Max pool 1	3 × 3	2	32×64×3232×64×32
Residual block 1	3 × 3	1	32×64×3232×64×32
Residual block 2	3 × 3	1	32×64×3232×64×32
Residual block 3	3 × 3	2	64×32×1664×32×16

Название слоя	Размер фильтра	Шаг	Количество выходов слоя
Residual block 4	3 × 3	1	64×32×1664×32×16
Residual block 5	3 × 3	2	128×16×8128×16×8
Residual block 6	3 × 3	1	128×16×8128×16×8
Dense layer 1		—	128
Batch norm		—	128

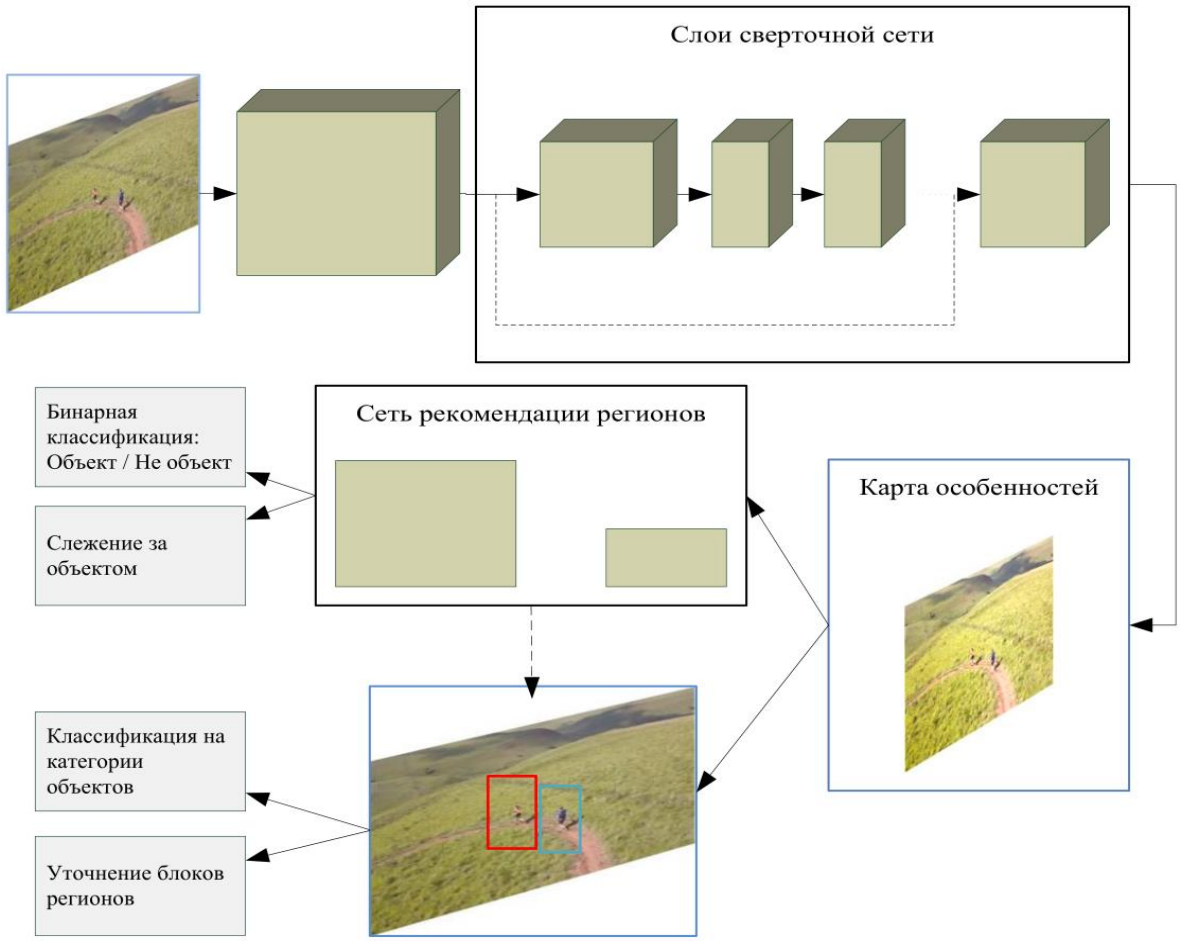

Рис. 1. Структура предложенной системы для отслеживания и классификации объектов
Fig. 1. The structure of the proposed object tracking system and objects classification

На первом этапе сеть предложения областей (RPN) генерирует привязки к регионам, присутствующим на изображении, которые имеют высокую вероятность присутствия объекта. Этот процесс делится на три этапа.

Первый этап включает в себя процесс извлечения признаков с использованием сверточной нейронной сети. Карты объектов свертки генерируются на последнем слое. На втором этапе используется подход скользящего окна на этих картах объектов для создания блоков, содержащих объекты. Параметры блоков уточняются на следующем этапе, чтобы указать на наличие объектов в них. Наконец, на третьем этапе сгенерированные маски уточняются с использованием более простой сети, которая вычисляет функцию потерь для выбора ключевых блоков, содержащих объекты. Для нейронной сети предложения областей обязательным шагом является извлечение функций свертки, которые вычисляются с использованием основной сети.

Экспериментальные исследования. Эффективность алгоритма отслеживания объектов оценивалась с использованием более 10 видеопоследовательностей из открытых баз данных: KITTI Vision Benchmark Suite [11], Drones Dataset [12], MOT16 [13], содержащих более 20000 кадров, для которых указаны границы объектов интереса: люди и машины. Сцены значительно различаются с точки зрения фона, условий освещения и способов движения камеры. Исследование позволяет определить точность обнаружения и отслеживания движущихся объектов в соответствии с известной метрикой Clear MOT, показывающей соотношение корректно обнаруженных пикселов на изображении, принадлежащих объектам интереса, к известному значению (Ground Truth) [14]:

$$\mathrm{MOTA} = 1 - \frac{\sum_t \left(FP_t + FN_t + ID_Sw_t \right)}{\sum_t GT_t},$$

где t – номер кадра видеопоследовательности; GT – достоверное значение, показывающее количество пикселов, содержащих объект интереса; FP и FN – соответственно ложные и ложноотрицательные срабатывания детектора; ID_Sw – порог смены идентичности объекта вследствие сложных траекторий и шума при наблюдении.

В табл. 2 приведены основные результаты работы системы слежения за объектами на различных видеопоследовательностях, а также параметры сцены. На рис. 2 показаны примеры отслеживания объектов.

Таблица 2

Оценка эффективности системы слежения

Наименование видеопоследовательности	**Разрешение**	**FPS**	**Количество кадров**	**Движение камеры**	**MOTA**
MOT16_07 [13]	1920×1080	30	500	Y	64,1
DJI_0574 [12]	3840×2160	60	960	Y	78,5
Bluemlisalphutte Flyover [12]	1280×720	60	990	Y	75,8
Berghouse Leopard Jog [12]	1280×720	30	1110	N	63,2
Kitty_0016 [11]	1920×1080	30	509	Y	58,1
Kitty_0018 [11]	1920×1080	60	179	Y	61,4
Kitty_0024 [11]	1920×1080	30	315	N	63,7

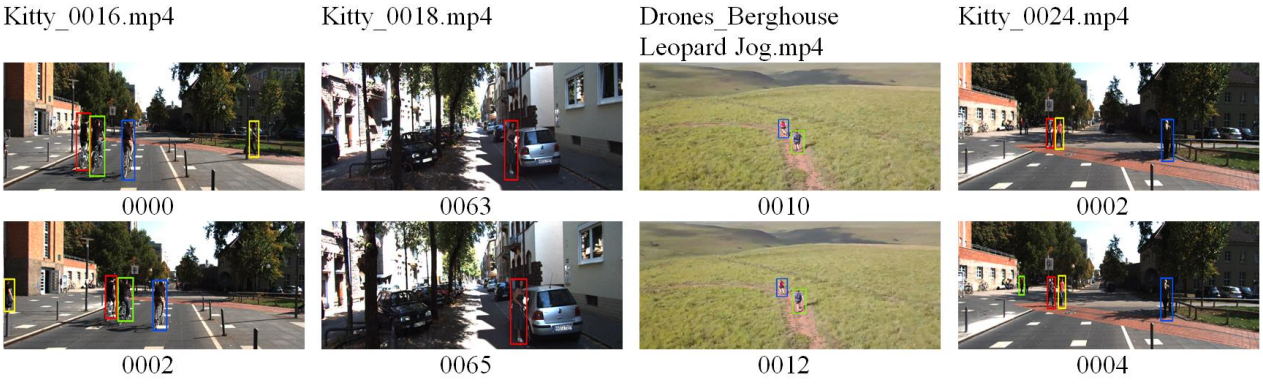

Рис. 2. Примеры отслеживания объектов тестовых видеопоследовательностей
Fig. 2. Tracking examples of testing videos

В табл. 3 представлена оценка основных параметров качества системы слежения за объектами в сравнении с другими современными системами.

Таблица 3

Оценка эффективности различных систем слежения за объектами

	MOTA	**MOTP**	**FP**	**FN**	**Runtime**
KNDT	68,2	79,4	11,479	45,605	0,7 Hz
POI	66,1	79,5	5061	55,914	10 Hz
SORT	59,8	79,6	8698	63,245	60 Hz
Deep Sort	61,4	79,1	12,852	56,668	40 Hz
Предложенная система	60,8	79,8	3855	37,45	42 Hz

Заключение. В результате разработана система, позволяющая осуществлять слежения за различными объектами для задач видеонаблюдения и видеоаналитики. Особенностями алгоритма является псевдореальная скорость работы: 25–35 кадров в секунду при разрешении 1920 × 1080, а также высокое качество слежения за объектами, что связано с применением нейронной сети для уточнения обнаруженных регионов. Для повышения удобства оператора и качества работы системы видеоаналитики система включает методы стабилизации видеопоследовательности, позволяющие улучшить как видеоматериал, получаемый в реальном времени, так и уже имеющееся видео путем устранения непреднамеренного дрожания.

References

1. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*. 2017, Vol. 60, No. 6, P. 84–90.
2. Socher R., Perelygin A., Jean Y. Wu, Chuang J., Manning C. D., Ng A. Y., Potts C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2013. P. 1631–1642.
3. Mnih V., Kavukcuoglu K., Silver D. et al. Human-level control through deep reinforcement learning. *Nature*. 2015, Vol. 518, P. 529–533. Doi: 10.1038/nature14236.
4. Khan G., Tariq Z., Khan M. Multi-Person Tracking Based on Faster R-CNN and Deep Appearance Features, Visual Object Tracking with Deep Neural Networks, Pier Luigi Mazzeo, Srinivasan Ramakrishnan and Paolo Spagnolo, IntechOpen. 2019. Doi: 10.5772/intechopen.85215.
5. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks Computer Vision and Pattern Recognition arXiv: 1506.01497, 04.01.2015.
6. Hui T.-W., Tang X., Loy C.-C. LiteFlowNet: A lightweight convolutional neural network for optical flow estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, 2018, P. 8981–8989.
7. Dosovitskiy A., Fischer P., Ilg E., Höusser P., Hazırbas C., Golkov V., van der Smagt P., Cremers D., Brox T. Flownet: Learning optical flow with convolutional networks. *IEEE International Conference on Computer Vision*, 2015, P. 2758–2766.
8. Wang M. et al. Deep Online Video Stabilization With Multi-Grid Warping Transformation Learning. *IEEE Transactions on Image Processing*, 2019, Vol. 28, No. 5, P. 2283–2292. Doi: 10.1109/TIP.2018.2884280.
9. Favorskaya M. N., Buryachenko V. V., Zotin A. G., Pahirka A.I. Video completion in digital stabilization task using pseudo-panoramic technique. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*. 2017, Vol. XLII-2/W4, P. 83–90.

10. Favorskaya M. N., Buryachenko V. V. Background extraction method for analysis of natural images captured by camera traps. *Informatsionno-upravliaiushchie sistemy*. 2018, No. 6, P. 35–45. Doi:10.31799/1684-8853-2018-6-35-45.

11. Geiger A., Lenz P., Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. *IEEE 2012 Conference on Computer Vision and Pattern Recognition, Providence, RI*, 2012, P. 3354–3361,

12. Drone Videos DJI Mavic Pro Footage in Switzerland Available at: https://www.kaggle.com/kmader/ /drone-videos (accessed 05.05.2019).

13. Milan A., Leal-Taixé L., Reid I., Roth S., Schindler K. MOT16: A Benchmark for Multi-Object Tracking. arXiv:1603.00831 [cs], (arXiv: 1603.00831), 2016.

14. Sun S., Akhtar N., Song H., Mian A. S., Shah M. Deep Affinity Network for Multiple Object Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, P. 1–15.

15. Zhou Xiangzeng, Xie Lei, Zhang Peng, Zhang Yanning. An Ensemble of Deep Neural Networks for Object Tracking. *IEEE International Conference on Image Processing, ICIP 2014*, 2014.

© Буряченко В. В., Пахирка А. И., 2020

Буряченко Владимир Викторович – кандидат технических наук, доцент; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: buryachenko@sibsau.ru.

Пахирка Андрей Иванович – кандидат технических наук, доцент; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: pahirka@sibsau.ru.

Buryachenko Vladimir Viktorovich – Cand. Sc., Associate Professor; Reshetnev Siberian State University of Science and Technology. E-mail: buryachenko@sibsau.ru.

Pahirka Andrei Ivanovich – Cand. Sc., Associate Professor; Reshetnev Siberian State University of Science and Technology. E-mail: pahirka@sibsau.ru.