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

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. A lot of tracking systems include two stages: 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 to track several objects in the frame, including those in the rough conditions of video surveillance. The paper provides an overview of modern methods for solving objects tracking tasks, among which the most promising one is deep learning neural networks application. The main approach used in this paper is neural networks for detecting regions (R-CNN), which has 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 classical in the field MOT(Multi-Object tracking) metric for objects tracking based on the known databases available in open sources. The effectiveness of the proposed system is compared to other well-known works.

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

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.

Introduction. In recent years, methods of measuring motion and tracking objects have achieved impressive results. Works on monitoring the movement of people in public places, facial recognition methods, suspicious objects detection and people's deviant behavior are practical applications for improving security. Video sequence stabilization methods are also widely used in video analytics as well as to improve the convenience of operators' work with surveillance systems. Stabilization eliminates unintentional video jitter, while preserving directed motions and camera panning.

Most of the tasks discussed are based on the use of high-level video sequence analysis, motion estimation, and object tracking techniques, which require demanding tasks to detect and track special points in video sequence frames. The application form proposes to use modern deep learning technologies to improve quality and reduce computational and time costs for performing motion estimation and stabilizing video sequences.

Review of publications. Deep neural networks have proved to be one of the best technologies for solving numerous problems related to digital image processing. The greatest success was demonstrated by convolutional neural networks containing from 10 to 300 layers, when solving image recognition problems [1], semantic analysis of texts [2] and training with reinforcement [3]. Recent studies have shown the effectiveness of artificial neural networks of complex structures when solving problems of motion analysis [4; 5] and evaluating optical flow [6; 7], as well as the possibility of using such technologies to stabilize video sequences [8]. One known approach used to detect and track objects is to estimate the visibility in the frame [9; 10].

This paper examines methods of tracking moving objects using deep learning approaches. To improve the quality of the algorithm, a convolutional network is used to classify regions (R-CNN) and methods for stabilizing the received video material. Visual tracking of objects is a classical computer vision task in which the position of the target is determined in each frame. This area of research remains in demand due to the large number of practical tasks based on tracking various objects. The algorithms used in this field are also improving, and allow to solve highly complex problems, such as occlusion, changing the position and shape of objects, people's appearance, lighting and presence of a complex background with various textures. In this regard, the publications offer a number of algorithms and approaches aimed at solving various tracking problems and improving the overall performance of object tracking.

Tracking of objects. A typical tracking system consists of two main models, the motion model and the appearance model. The motion model is used to predict the target location in the subsequent frame, similarly to the use of the Kalman filter or particle filter to simulate the target motion. The motion model can also be simple, for example including linear motion, on the basis of which several more complex trajectories are built; and more complex, which track objects taking with respect to changes in direction and speed of movement. To speed up the motion evaluation process, a motion value assumption is proposed within the search box around the previous location of the object. On the other hand, the appearance model is used to describe the target and check the predicted location of the target in each frame. Appearance models can use generative and discriminatory methods. In generative methods, tracking is performed by searching for a region most similar to an object. Discriminatory methods use a classifier that allows to distinguish between the object and the background. In general, the appearance model can be updated during system operation, taking into account required changes to objects. This allows, for example, to continue tracking a person when turning or tilting a body.

Traditionally, motion tracking algorithms have used manually calculated functions based on pixel intensity, color, and histogram of oriented gradients (HOG) to represent the target in generative or discriminatory description models. Although they achieve satisfactory performance under certain conditions, they are not resistant to major changes in the appearance of objects. Deep training using Convolution Neural Networks (CNN) has recently significantly improved the performance of various computer vision applications.

This approach also affected visual tracking of objects and partially allowed to overcome difficulties and get better performance compared to the methods used earlier. In CNN-based tracking systems, the object appearance model is based on convolutional network training, and the classifier is used to mark the path on the image as belonging to the object or background. CNN-based systems have achieved a modern level of efficiency even using simple off-line motion models without retraining. However, such systems usually experience high computational loads due to the large number of possible trajectories of objects during the stages of neural network training and objects tracking.

A promising trend in the field of object tracking is the tasks associated with the analysis of a large number of people in a frame (Multi-object tracking (MOT)). This task has two stages: detecting objects and associating them in different frames. During the first stage, desired objects are detected in each frame of the video stream, where the objects may be different depending on the detector used. The quality of detection directly affects the performance of the tracking system. The second stage includes searching for a match between detected objects in the current frame and the previous one to estimate their motion paths. The high accuracy of the object detection system results in fewer missing objects and more stable trajectories. However, at the same time, excessive retraining of the detection system can reduce the quality of the system when changing object parameters. Applying additional approaches when solving the problem of associating objects in different frames can improve the quality of tracking complex objects with changing parameters. The accuracy of object detection can be increased by using a convolutional neural network based on deep learning. Objects are merged based on appearance and improved motion functions.

R-CNN-based object tracking system. One of the areas of research is the development of an algorithm that allows to improve tracking objects accuracy by using a convolutional network for classifying regions (R-CNN) and increasing the speed of the system to close to real time evaluation. A convolutional network allows to classify an area as belonging to an object or background, and at the same time, the characteristic map obtained during the network operation is also used to perform approximate localization of objects and allows to reduce the time for evaluating matches between them.

Practical experiments were conducted applying the neural network to classify regions, the structure of which is shown in tab. 1. The main task is to track several types of objects in a frame: people and cars. Proposed method includes components of human detection, prediction of objects position in subsequent frames, association of detected objects and control of selected objects cycle. The most widely known object detection algorithm is YOLO [24], which is fast enough to detect several objects in real time, but has insufficient accuracy, leading to trajectories fragmentation and complexities with identifying detected objects.

The basic diagram of the developed system is shown in fig. 1. With the development of deep learning-based algorithms, detection of objects in complex conditions has become much easier. A key component of the algorithm is a convolutional region detection network (R-CNN).

During the first stage, a region proposal network (RPN) generates bindings to regions in the image that have a high probability of an object presence. This process is divided into three steps.

Table 1

Convolutional region network parameters

Layer name	Filter dimensions	Step	Number of layer outputs
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
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

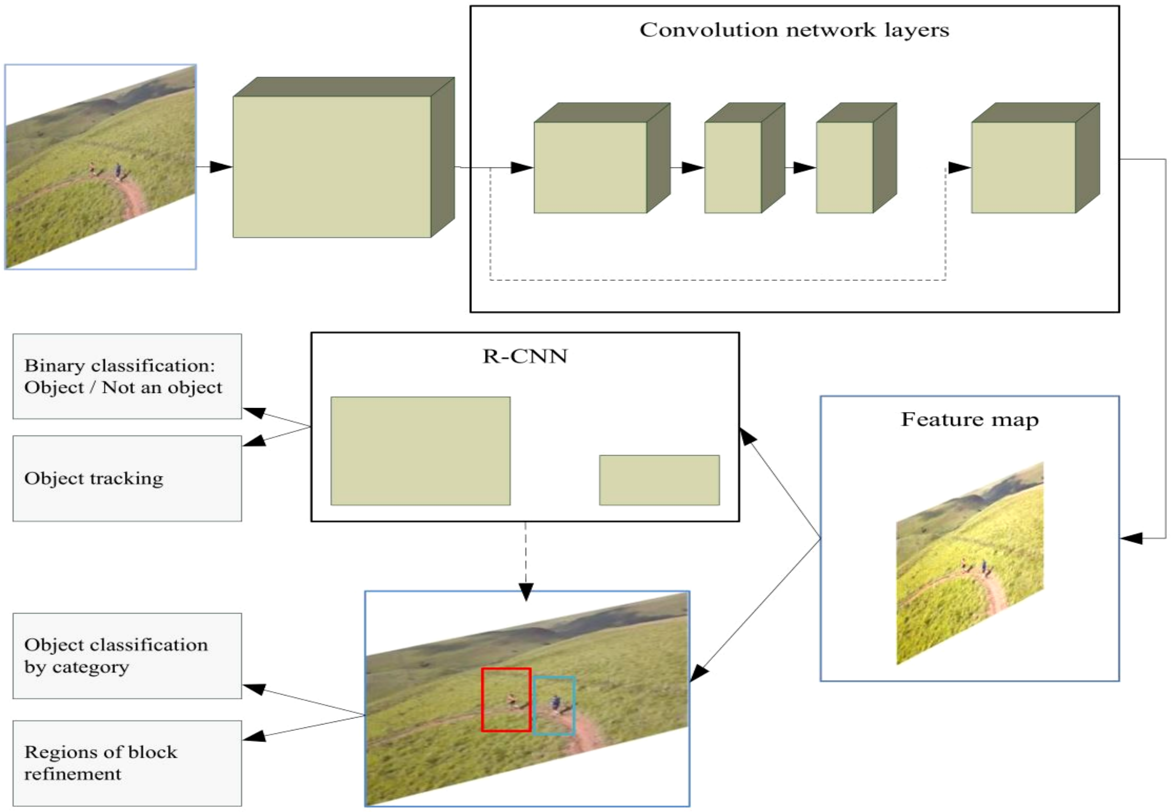

Fig. 1. The structure of the proposed object tracking system and objects classification

Рис. 1. Структура предложенной системы для отслеживания и классификации объектов

The first one includes a feature extraction process using a convolutional neural network. Convolution object maps are generated on the last layer. The second step uses the sliding window approach on these object maps to create blocks containing objects. The block parameters are refined in the next step to indicate the presence of objects in them.

Finally, in the third step, the generated masks are refined using a simpler network that calculates a loss function for selecting key blocks containing objects. For a neural network, proposing regions is a necessary step in extracting convolution functions that are calculated using the main network.

Pilot studies. Effectiveness of the object tracking algorithm was assessed using more than 10 video sequences from open databases: KITTI Vision Benchmark Suite [11], Drones Dataset [12], MOT16 [13], containing more than 20,000 frames for which the boundaries of objects of interest were indicated: people and machines. Scenes vary significantly in terms of background, lighting conditions, and how the camera moves. The study allows to determine the accuracy of detecting and tracking moving objects in accordance with the known metric Clear MOT (1), showing the ratio of correctly detected pixels in the image belonging to objects of interest to a known value (Ground Truth) [14]:

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

where t – the frame number of the video sequence; GT – a valid value indicating the number of pixels containing the object of interest; FP and FN are false positive and false negative detectors respectively; ID_Sw is the threshold for changing object identity due to complex trajectories and noise during observation.

Tab. 2 shows the main results of object tracking system operation on various video sequences, as well as the scene parameters. Fig. 2 shows examples of object tracking.

Tab. 3 presents the assessment of the basic object tracking system quality parameters in comparison with other modern systems.

Table 2

Tracking system efficiency assessment

Video sequence name	Resolution	FPS	Number of frames	Camera motion	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

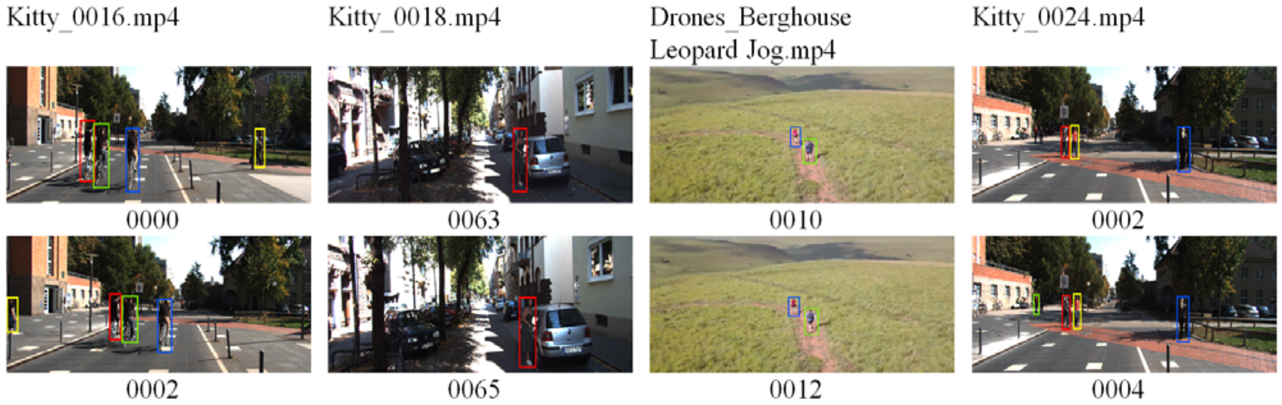

Fig. 2. Object racking examples of testing videos

Рис. 2. Примеры отслеживания объектов тестовых видеопоследовательностей

Table 3

Assessment of different object tracking systems efficiency

	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
Proposed system	60.8	79.8	3855	37.45	42 Hz

Conclusion. As a result, the system has been developed to track various objects for video surveillance and video analytics. Features of the algorithm are pseudo-real speed: 25–35 frames per second at the resolution of 1920 × 1080, as well as high quality of object tracking, which is associated with the use of the neural network to clarify the detected regions. To improve operator convenience and video analysis system quality, the system includes video sequence stabilization techniques that improve both real-time video and existing video by eliminating unintentional jitter.

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., Hösser 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.

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.

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.