This is a preprint of a conference proceedings published in IEEE 9th International Workshop on Metrology for AeroSpace (MetroAeroSpace), (2022) 270-274
The final authenticated version is available online at: https://doi:10.1109/MetroAeroSpace54187.2022.9855900

Robust S-Y-biLSTM object tracking method for on-road objects shoot from an unmanned aerial vehicle

1st Ivan Saetchnikov
Radiophysics Department
Belarusian State University
Minsk, Belarus
saetchnikov@bsu.by

2nd Victor Skakun
Radiophysics Department
Belarusian State University
Minsk, Belarus
skakun@bsu.by

3rd Elina Tcherniavskaia
Physics Department
Belarusian State University
Minsk, Belarus
tcherniavskaia@bsu.by

***Abstract*— being one of the most challenging tasks in computer vision dynamical object tracking problems have faced additional issues regarding the unmanned aerial vehicle. In particular, image degradation, uneven object intensity, variety in object sizes, etc. In this paper, we proposed the implementation of the S-Y-biLSTM technique for on-road object shooting from an unmanned aerial vehicle. The training and testing process of the proposed network was performed based on the dataset of on-road objects shouted at the height of 15-45 meters.**

Keywords*— multi-object tracking, deep neural network, convolutional neural network, recurrent neural network, object detection.

I. Introduction

Over the past several decades the computer vision area is moving highly forward in the range of tasks, e.g., motion analysis [1], object capture [1], object detection [3], traction, etc. primarily due to the opportunities provided by the deep learning-based methods. Its high speediness and accuracy result in feature extraction and scaling has extended the range of applications where such systems may be utilized. [4] One of the most challenging and high-promising domain application area is the unmanned aerial vehicle. [5 – 12]

It is worth to mention, that appropriate image processing technique of single-shot camera-based computer vision system in the long-term perspectives seems one of the key competitive techniques in such system due to the high cost of utilizing and deployment of alternative systems. For example, sensing systems with Radiofrequency (RF) ranging [13], LIDAR-based systems [14] or other multisensory systems [15]. To build and validate the deep-learning based algorithms fairly, the data science community has successfully developed a number of datasets including detection datasets (e.g., Caltech [16] and DETRAC [16]) and tracking datasets (e.g., KITTI-T [17] and VOT2016 [18]). The visual tracking approach using deep learning neural networks has been not only significantly improved compared to traditional approaches [19], but also improved over the last 10 years in terms of deep learning approaches [20].

In [21] the authors presented the DeepSort-based network that used the pre-trained convolutional neural network to compute bounding boxes. The authors of [22] proposing the Tracktor ++ network have progressed in accurate and speed by using a Faster R-CNN [23] for frame-to-frame tracking by extracting features. Other suggestions, for example, Joint Detection and Embedding (JDE) [24] network based on RetinaNet [25] architecture, deviated from the two-stage paradigm.

However, despite the high promising perspectives of the computer vision-based systems, the object detection and tracking scenarios from the unmanned aerial vehicle has accompanied by the range of additional challenges [26 – 30]:

- Image degradation. Rapid changes in the movement within the external environment cause noisy and fuzzy aerial images. In addition, high-speed flight or camera rotation also increases the complexity of object detection.
- Uneven object intensity. Flexibility in camera movement may result in an uneven density of captured and detected objects. [16] Also, most objects occupy a small part of the entire image, that resulting in challenges to separate them from their surrounding environment.
- Real-time problem. Object tracking is highly different from the detection or classification task due to the requirement to accurately locate the observing objects in real time.

In this paper we proposed an implementation of proposed S-Y-biLSTM network for object tracking from unmanned aerial vehicle. According to our hypothesis, the memory - based methods will be highly effective for the task of object tracking via the unmanned aerial vehicle, because on-road objects regularly move along a strictly trajectories from the UAV's shooting view. To validate our idea, we have collected the dataset based on image sequences shooted from the unmanned aerial vehicles shouted: VisDrone 2019 [31], Drone Vehicle Dataset [32], and DTB70 [33] with the corresponding labels. Finally, we have provided the robustness analysis of S-Y-biLST compared to LYOLOv4eff [34], ROLO [35] and DeepSort. [32]

II. R-S-biLSTM object tracking method

In this paper we have presented the R-S-biLSTM object tracking method for on-road objects shoot from an unmanned aerial vehicle consisted of 3 major steps. The key motivation

This is a preprint of a conference proceedings published in IEEE 9th International Workshop on Metrology for AeroSpace (MetroAeroSpace), (2022) 270-274
The final authenticated version is available online at: https://doi:10.1109/MetroAeroSpace54187.2022.9855900

behind the proposed method is that for the tracking technique, the learning process based on historical visual information seems like an optional approach. In our vision, within the tracking technique the recurrent convolutional-based network will be doubled the effectiveness due to the fact, that it uses both the history of locations and robust visual features of past frames. In particular, the Long Short Time Memory (LSTM) [36 – 41] using memory cells better discover the long-range temporal relations. This method extends the neural network learning and analysis into the spatial and temporal domain.

Firstly, we recalculated the difference between the following frames and passed it through the YOLO-based network to extract the feature map. All skip connection outputs, and the final output, are concatenated together and fed through a final fully connected layer to further reduce the dimensionality of the embedding space. The YOLO-based features based on our hypothesis will be more robust for using as an input of LSTM network. As the detection technique, we proposed the SSD_eff method. The bidirectional LSTM was used as a backbone tracking technique due to possibility to remember longer term relationship. Compared to the backbone LSTM network [45], where spatial information is lost due to the encoded input, in our method we have replaced the fully connected layers with the convolutional ones. Thus, in our case in the input-to-state and state-to-state transitions the fully spatial-temporal correlation information is used.

The detailed equations' formulation of biLSTM sell is presented in the 1 – 5 below:

$$i(t) = \text{sigma}\ (W(x,i) * x(t)+W(h,i) * H(t-1)+b(i)); \quad (1)$$

$$k(t) = \text{sigma}\ (W(x,k) * x(t)+W(h,k) * H(t-1)+b(k)); \quad (2)$$

$$f(t) = \text{sigma}\ (W(x,f) * x(t)+W(h,f) * H(t-1)+b(f)); \quad (3)$$

$$CELL(t) = f(t) * CELL(t-1) + i(t) * \tanh\ (W(x) * X(t) + W(h) * H(t-1) + b(cell)); \quad (4)$$

$$H(t) = k(t) * \tanh\ (CELL(t)) \quad (5)$$

In equations (1) – (5) "*" – convolution operator, "x" – Hadamard product, σ is the sigmoid function and W(x*), W(h*) - convolutional kernels (input and hidden respectively). The hidden H(t) and cell states C(t) updated based on the input states X(t) that pass via i(t), f(t), k(t) gates activations during time steps, b – bias terms.

The bidirectional LSTM can thereby access long-range context in both directions of the time sequence of the input, therefore potentially gain a better understanding of the video sequence.

To sum up, the set (h(f), c(f)) from the first LSTM cell is redirected on the forward pass, the second set (h(b), c(b)) on the backward pass. In additional it is worth to mention, that for each time sequence the corresponding hidden states from the two LSTM sets after stacking are passing through a Convolution layer to build a resulted hidden representation, that will be redirected to the next time step.

III. Results

We have analyzed datasets with videos of objects in the cities shooting from UAV's on-board cameras. Finally, our dataset was collected our based-on parts of three datasets: VisDrone 2019 [31], Drone Vehicle Dataset [32], and DTB70 [33] with included pre-labeled annotations of bounding boxes. In our task, we collected sequences of objects in side-shot scope mode from the height level (15 – 45 meters). Within the preprocessing stage, we have resized images to a single resolution 1024x540 using the pre-trained ImresNet [46] network based on residual learning strategy. The collected dataset includes 300 video sequences 3.4 hours long with the corresponding annotations of objects. The average area occupied by detected and tracked objects was about 3,4% pixels of the whole frame, the maximum was 23,6%. Our method we have compared to the up-to-date tracking methods: ROLO [47], DeepSort [48] and LYOLOv4eff [35]. First, it is worth to mention, that the proposed method surpassed the LYOLOv4eff, DeepSort and ROLO based on the majority of MOT metrics presented in the Table 1. In particular, we have dramatically decreased the number of ID switches in contrast to LYOLOv4eff, that with the lower retention level by MT metric demonstrate the significant positive shift in the tracking robustness in compared to LYOLOv4eff. However, we have mentioned tracked objects during less than 20% of its lifespan by the proposed network. However, higher value of ML metric leveled out by the higher values by FN and FP metrics. To sum up, considering the 2 major MOT metrics MOTA and MOTP we have mentioned that the proposed method achieved the higher robustness and accuracy ability in compared to LYOLOv4eff [34] ROLO and DeepSort.

IV. Conclusion

With this paper, we have proposed a novel S-Y-biLSTM network for dynamical on-road object tracking from the aerial vehicle, in particular, unmanned aerial vehicles. (UAV) It consisted of 3 major stages. The performance analysis of the proposed approach has shown that our network S-Y-biLSTM surpasses not only the LYOLOv4eff but also ROLO and DeepSort methods based on 2 cumulative MOT metrics MOTA, MOTP with the following results for S-Y-biLSTM compared to LYOLOv4eff, DeepSort and ROLO.

References

[1] W. Wei and A. Yunxiao, "Vision-Based Human Motion Recognition: A Survey," 2009 Second International Conference on Intelligent Networks and Intelligent Systems, 2009, pp. 386-389, doi: 10.1109/ICINIS.2009.105.

[2] H. Liao, P. Chen, Z. Lin and Z. Lim, "Automatic zooming mechanism for capturing clear moving object image using high definition fixed camera," 2017 19th International Conference on Advanced Communication Technology (ICACT), 2017, pp. 869-876, doi: 10.23919/ICACT.2017.7890238.

[3] F. Ajmera, S. Meshram, S. Nemade and V. Gaikwad, "Survey on Object Detection in Aerial Imagery," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1050-1055, doi: 10.1109/ICICV50876.2021.9388517.

[4] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu and N. Yu, "Online Multi-object Tracking Using CNN-Based Single Object Tracker with Spatial-Temporal Attention Mechanism," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 4846-4855, doi: 10.1109/ICCV.2017.518.

This is a preprint of a conference proceedings published in IEEE 9th International Workshop on Metrology for AeroSpace (MetroAeroSpace), (2022) 270-274
The final authenticated version is available online at: https://doi:10.1109/MetroAeroSpace54187.2022.9855900

[5] Bochkovskiy, Alexey & Wang, Chien-Yao & Liao, Hong-yuan. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection.

[6] M. -R. Hsieh, Y. -L. Lin and W. H. Hsu, "Drone-Based Object Counting by Spatially Regularized Regional Proposal Network," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4165-4173, doi: 10.1109/ICCV.2017.446.

[7] I. Guvenc, F. Koohifar, S. Singh, M. L. Sichitiu and D. Matolak, "Detection, Tracking, and Interdiction for Amateur Drones," in IEEE Communications Magazine, vol. 56, no. 4, pp. 75-81, April 2018, doi: 10.1109/MCOM.2018.1700455.

[8] Z. Huang, C. Fu, Y. Li, F. Lin and P. Lu, "Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2891-2900, doi: 10.1109/ICCV.2019.00298.

[9] P. Zhang, Y. Zhong and X. Li, "SlimYOLOv3: Narrower, Faster and Better for Real-Time UAV Applications," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 37-45, doi: 10.1109/ICCVW.2019.00011.

[10] Y. Wu, Y. Sui and G. Wang, "Vision-Based Real-Time Aerial Object Localization and Tracking for UAV Sensing System," in IEEE Access, vol. 5, pp. 23969-23978, 2017, doi: 10.1109/ACCESS.2017.2764419.

[11] Y. Li, C. Fu, F. Ding, Z. Huang and G. Lu, "AutoTrack: Towards High-Performance Visual Tracking for UAV With Automatic Spatio-Temporal Regularization," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11920-11929, doi: 10.1109/CVPR42600.2020.01194.

[12] Y. Wu, J. Lim and M. Yang, "Online Object Tracking: A Benchmark," 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2411-2418, doi: 10.1109/CVPR.2013.312.

[13] D. Du et al., "VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 213-226, doi: 10.1109/ICCVW.2019.00030.

[14] Du, Dawei & Qi, Yuankai & Yu, Hongyang & Yang, Yifan & Duan, Kaiwen & Li, Guorong & Zhang, Weigang & Tian, Qi. (2018). The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking.

[15] P. Bergmann, T. Meinhardt and L. Leal-Taixé, "Tracking Without Bells and Whistles," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 941-951, doi: 10.1109/ICCV.2019.00103.

[16] Belmonte, L. & Morales, R. & Fernández-Caballero, Antonio. (2019). Computer Vision in Autonomous Unmanned Aerial Vehicles—A Systematic Mapping Study. Applied Sciences. 9. 3196. 10.3390/app9153196.

[17] E. Martinez-Martin and A. P. del Pobil, "Object Detection and Recognition for Assistive Robots: Experimentation and Implementation," in IEEE Robotics & Automation Magazine, vol. 24, no. 3, pp. 123-138, Sept. 2017, doi: 10.1109/MRA.2016.2615329.

[18] P. Voigtlaender et al., "MOTS: Multi-Object Tracking and Segmentation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 7934-7943, doi: 10.1109/CVPR.2019.00813.

[19] L. Jiao, D. Wang, Y. Bai, P. Chen and F. Liu, "Deep Learning in Visual Tracking: A Review," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3136907.

[20] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan and S. Kasaei, "Deep Learning for Visual Tracking: A Comprehensive Survey," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 5, pp. 3943-3968, May 2022, doi: 10.1109/TITS.2020.3046478.

[21] Shivani Kapania, etc. 2020. Multi Object Tracking with UAVs using Deep SORT and YOLOv3 RetinaNet Detection Framework. In Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems (AIMS '20). Association for Computing Machinery, New York, NY, USA, Article 1, 1–6. DOI:https://doi.org/10.1145/3377283.337

[22] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[23] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[24] Q. Liu, B. Liu, Y. Wu, W. Li and N. Yu, "Real-Time Online Multi-Object Tracking in Compressed Domain," in IEEE Access, vol. 7, pp. 76489-76499, 2019, doi: 10.1109/ACCESS.2019.2921975.

[25] Z. Tong, L. Jieyu and D. Zhiqiang, "UAV Target Detection based on RetinaNet," 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 2019, pp. 3342-3346, doi: 10.1109/CCDC.2019.8832517.

[26] W. Yu, T. Yang and C. Chen, "Towards Resolving the Challenge of Long-tail Distribution in UAV Images for Object Detection," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 3257-3266, doi: 10.1109/WACV48630.2021.00330.

[27] T. Nakamura, S. Haviland, D. Bershadsky, D. Magree and E. N. Johnson, "Vision-based closed-loop tracking using micro air vehicles," 2016 IEEE Aerospace Conference, 2016, pp. 1-12, doi: 10.1109/AERO.2016.7500873.

[28] A. Koksal, K. G. Ince and A. Aydin Alatan, "Effect of Annotation Errors on Drone Detection with YOLOv3," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 4439-4447, doi: 10.1109/CVPRW50498.2020.00523.

[29] M. Hossain, M. A. Hossain and F. A. Sunny, "A UAV-Based Traffic Monitoring System for Smart Cities," 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI), 2019, pp. 1-6, doi: 10.1109/STI47673.2019.9068088.

[30] R. Tao, E. Gavves and A. W. M. Smeulders, "Siamese Instance Search for Tracking," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1420-1429, doi: 10.1109/CVPR.2016.158.

[31] S. Desai and D. Ghose, "Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1485-1495, doi: 10.1109/WACV51458.2022.00155.

[32] R. Krajewski, J. Bock, L. Kloeker and L. Eckstein, "The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 2118-2125, doi: 10.1109/ITSC.2018.8569552.

[33] R. Sun, L. Fang, X. Gao and J. Gao, "A Novel Target-Aware Dual Matching and Compensatory Segmentation Tracker for Aerial Videos," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-13, 2021, Art no. 5015613, doi: 10.1109/TIM.2021.3109722.

[34] I. Saetchnikov, V. Skakun and E. Tcherniavskaia, "Efficient objects tracking from an unmanned aerial vehicle," 2021 IEEE 8th International Workshop on Metrology for AeroSpace (MetroAeroSpace), 2021, pp. 221-225, doi: 10.1109/MetroAeroSpace51421.2021.9511748.

[35] G. Ning et al., "Spatially supervised recurrent convolutional neural networks for visual object tracking," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), 2017, pp. 1-4, doi: 10.1109/ISCAS.2017.8050867.

[36] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink and J. Schmidhuber, "LSTM: A Search Space Odyssey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222-2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.

[37] Y. Tian and L. Pan, "Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network," 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), 2015, pp. 153-158, doi: 10.1109/SmartCity.2015.63.

[38] L. Mou, P. Ghamisi and X. X. Zhu, "Deep Recurrent Neural Networks for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 3639-3655, July 2017, doi: 10.1109/TGRS.2016.2636241.

[39] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in IEEE Access, vol. 6, pp. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

This is a preprint of a conference proceedings published in IEEE 9th International Workshop on Metrology for AeroSpace (MetroAeroSpace), (2022) 270-274
The final authenticated version is available online at: https://doi:10.1109/MetroAeroSpace54187.2022.9855900

[40] V. Veeriah, N. Zhuang and G. Qi, "Differential Recurrent Neural Networks for Action Recognition," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4041-4049, doi: 10.1109/ICCV.2015.460.

[41] Shaohua Kevin Zhou, R. Chellappa and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," in IEEE Transactions on Image Processing, vol. 13, no. 11, pp. 1491-1506, Nov. 2004, doi: 10.1109/TIP.2004.836152.

[42] I. Saetchnikov, E. A. Tcherniavskaia and V. V. Skakun, "Object detection for unmanned aerial vehicle camera via convolutional neural networks," in IEEE Journal on Miniaturization for Air and Space Systems, doi: 10.1109/JMASS.2020.3040976.

[43] Y. Wang, C. Wang and H. Zhang, "Combining single shot multibox detector with transfer learning for ship detection using Sentinel-1 images," 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), 2017, pp. 1-4, doi: 10.1109/BIGSARDATA.2017.8124924.

[44] C. Alippi, S. Disabato and M. Roveri, "Moving Convolutional Neural Networks to Embedded Systems: The AlexNet and VGG-16 Case," 2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), 2018, pp. 212-223, doi: 10.1109/IPSN.2018.00049.

[45] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[46] I. Saetchnikov, V. Skakun and E. Tcherniavskaia, "Pattern recognition on aerospace images using deep neural networks," 2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace), Pisa, Italy, 2020, pp. 336-340, doi: 10.1109/MetroAeroSpace48742.2020.9160198.

[47] Ning, Guanghan & Zhang, Zhi & Huang, Chen & Ren, Xiaobo & Wang, Haohong & Cai, Canhui & He, Zhihai. (2017). Spatially supervised recurrent convolutional neural networks for visual object tracking. 1-4. 10.1109/ISCAS.2017.8050867.

[48] Y. Gai, W. He and Z. Zhou, "Pedestrian Target Tracking Based On DeepSORT With YOLOv5," 2021 2nd International Conference on Computer Engineering and Intelligent Control (ICCEIC), 2021, pp. 1-5, doi: 10.1109/ICCEIC54227.2021.00008.