A One-stage Temporal Detector with Attentional LSTM for Video Object Detection

Jiahui Yu , Zhaojie Ju , Hongwei Gao and Dalin Zhou

***Abstract*— Temporal object detection is more challenging than static image detection because of the rich context information. Recently, state-of-the-art works mine context information to detect each frame by using LSTM-based modules. However, restricted by the low-exploration of temporal information, significant results in terms of accuracies and speeds are not reported by the existing methods. In this paper, we propose a new one-stage temporal detector for online video object detection. A new structure with an improved spatiotemporal LSTM (STLSTM) is proposed to suppress useless background information. Next, the SSD-based structure is improved to extract rich features and high-level semantic features. We evaluate the proposed model on the ImageNet benchmark and space human-robot interaction database. Extensive comparisons show that the proposed detector achieves state-of-the-art performance.**

I. INTRODUCTION

As a critical application of human-robot interaction, video object detection has received increasing interest. Rich temporal information mining is the key to improve the detection performance in videos. Recently, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN) have been extended to exploit spatiotemporal features in sequence learning tasks.

A. Related work

State-of-the-art models have been proposed for object detection based on static images, such as [1], [2], [3], and [4]. However, most of these works cannot achieve satisfactory results in videos. This is because they cannot exploit the rich temporal information of video sequences. Recently, many state-of-the-art detectors have been proposed to learn spatiotemporal features effectively, effective for video object detection. typical models are [5], [6] , [7], [8], [9], [10], and [11]. These works are divided into two categories according to the backbone type, including one-stage detectors and two-stage detectors. For example, most two-stage detectors are designed based on the ResNet with the region proposal networks (RPNs). Normally, this model can achieve higher recognition accuracy because its deeper backbone can obtain high-performance semantic features. In most real-world applications, however, one-stage detectors have been widely applied because of faster speeds. Motivated by the success of the RNNs and attention models, some state-of-the-art models provide some useful solutions on video sequence analysis tasks, such as [9], [10], [5], [12], and [13].

However, these existing video analysis methods yield different limitations because of various reasons. The utilization of the temporal features between frames is very low. Although the LSTM-based networks are adequate for mining temporal features with attention mechanisms, few studies have reported successfully modelling intra-frame and inter-frame temporal information simultaneously.

B. Motivation and contributions

In this paper, we proposed a one-stage temporal detector based on the SSD and the LSTM. A structure of multi-level feature modelling is proposed to take full use of temporal features in inter-frame and intra-frame. An improved spatiotemporal LSTM (ST-LSTM) is employed to mining rich context information. This is a new structure, which proves to suppress useless background information. Therefore, the proposed model achieves a significant improvement of detection accuracies and speeds in video object tasks.

The contributions of this work are summarized as follows.

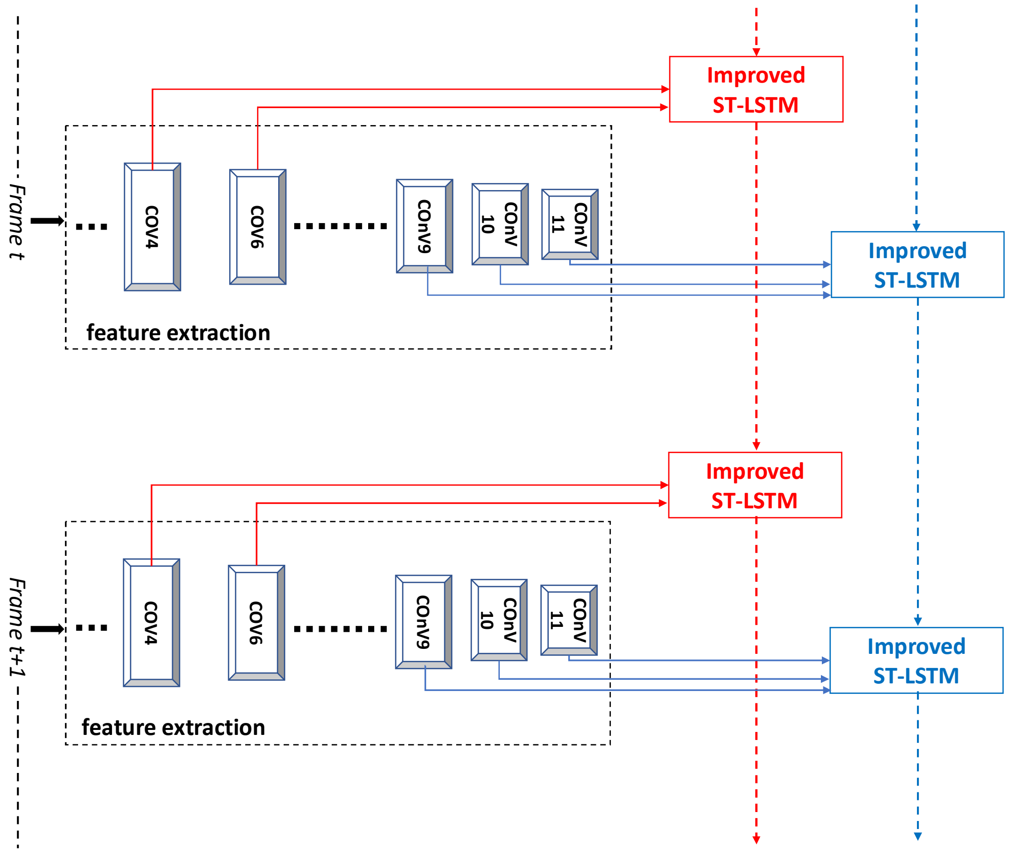
- A new structure is proposed to mine the temporal information in videos fully. Both inter-frame and intra-frame can be exploited effectively across time.
- To suppress useless information, an improved ST-LSTM is proposed. Attention maps flow between frames to enhance the temporal relationship of the context.
- The significant results are achieved on the ImageNet benchmark and a space human-robot interaction (SHRI) dataset in terms of detection accuracies and speeds.

II. PROPOSED TEMPORAL DETECTOR

In this section, we first describe the structure of the proposed model, including an SSD-based network for feature extraction, an LSTM-based network for context information analysis, and a memory enhancement module. Next, the technical details about the proposed memory enhancement method are described. Finally, we present an association training scheme in detail.

A. Backbone

The structure is shown in Fig. 1. Extending from the SSD as the feature extraction operator, a temporal structure is designed. For each frame, the shallow layers of the SSD are effective for small-object detection because richer features are included. The deep layers of the SSD are utilized to mine high-performance semantic features. Specifically, the



features of the $Conv4$ and $Conv6$ are inputted into an ST-LSTM to extract location features, and that of the $Conv9$, $Conv10$, and $Conv11$ are modelled in an ST-LSTM for high-level semantic features extraction. Note that the category-discriminative scores are used to help quickly generate candidate boxes at each frame, which is inspired by the TSSD [5]. This is faster than other works that utilize the RPNs. After that, both location features and semantic features are the input of the memory enhancement module. The spatial resolution of the layers are 300 × 300 (input image), 38 × 38 × 512 ($Conv4$), 19 × 19 × 512 ($Conv6$), 5 × 5 × 256 ($Conv9$), 3 × 3 × 256 ($Conv10$), and 1 × 1 × 256 ($Conv11$).

TABLE I

PERFORMANCE OF VARIOUS DESIGNS

NO.	Methods	mAP (%)
1	SSD	39.2
2	Improved SSD	48.3
3	LSTM-based	57.3
5	**Proposed**	**64.1**

B. *Temporal feature learning*

To suppress useless information, including background information and confusion information, first, a group of LSTM-based networks is designed. Next, compared to image object detection, the temporal information between each frame should be mined effectively in video object detection. Both the same structures are utilized to model features obtained from the SSD. According to the hierarchical relationship, these features are divided into two categories, including the shallow features and the deep semantic features. The pyramidal feature network motivates this. As shown in Fig. 1, the LSTM contained the elements of the $Conv4$ and $Conv6$ is denoted as low-level temporally operator, and that of $Conv9$, $Conv10$, and $Conv11$ is denoted as high-level temporally operator. After that, both location features and semantic features are obtained without useless temporal information.

The operator that we use to mine the relationship between frames is the ST-LSTM proposed in [14], which is partly motivated by the ConvLSTM [15]. Based on this, two improvements are conducted: 1) The attention maps are used to transform between frames rather that the features of hidden layers; 2) A feature preprocessing module is designed to change the dimension of the input feature that consists of three convolution layers. Technically, the improved ST-LSTM is computed as (1). Where a_t^{ST} is the attention map, and χ_t is the raw input; m_t^{k-1}, h_{t-1}^k, and c_{t-1}^k are the memory status; i_t , g_t , f_t, and o_t are gates; $w_{1\times1}$ denotes

TABLE II

COMPARISON OF THE PROPOSED MODEL AND OTHER STATE-OF-THE-ART METHODS ON THE IMAGENET VID DATASET

Methods	Based network	Training	Real-time	mAP (%)
Closed-loop [16]	VGG	Online	No	50
Seq-NMS [17]	VGG	Online	No	52.2
Bottleneck-LSTM [18]	MobileNet	Online	Yes	54.4
STMN [19]	VGG	Offline	No	55.6
TCNN [20]	Craft	Online	No	64.5
TSSD-OTA [21]	VGG	Online	Yes	65.4
Proposed	VGG	**Online**	**Yes**	**64.1**

the convolution layer, and σ denotes the sigmoid function.

$$\begin{aligned}
&a_t^{ST} = \sigma(w_i * [\chi_t, h_{t-1}]) \otimes \chi_t \\
&i_t = \sigma(w_{ia} * a_t^{ST} + w_{ih} * h_{t-1} + b_i) \\
&g_t = \tanh(w_{ga} * a_t^{ST} + w_{gh} * h_{t-1} + b_g) \\
&f_t = \sigma(w_{fa} * a_t^{ST} + w_{fh} * h_{t-1} + b_f) \\
&c_t^k = i_t \odot g_t + f_t \odot c_t^{k-1} \\
&m_t^k = {i_t}' \odot {g_t}' + {f'}_t \odot m_t^{k-1} \\
&o_t = \sigma(w_{oa} * a_t^{ST} + w_{oh} * h_{t-1} + w_{om} m_t^k + b_o) \\
&h_t^k = o_t \odot \tanh(w_{1\times 1} * [c_t^k, m_t^k])
\end{aligned}$$

III. EXPERIMENT

A. Experimental settings

Database. We evaluate the proposed method on two datasets, including the ImageNet VID benchmark [22], and the SRSSL dataset [23]. The former is a large-scale benchmark for video object detection and also is a huge challenge. The latter is a dataset collected in 2020 for the Space human-robot interaction (SHRI) task consisting of 6000 samples. Eight hand gestures for the SHRI task are included, and multi-size objects are collected to show the SHRI scene. We follow the protocols given in [24], [23], and [25] to evaluate the proposed method and utilize the mean average precision (mAP) as the evaluation index.

TABLE III

MPAS ON DIFFERENT SIZE OBJECTS ON THE SRSSL DATASET

Method	Size				
	XS	S	M	L	XL
SSD-300	62.38	93.75	89.43	90.5	93.48
DSSD	88.47	87.74	88.53	92.32	92.43
FF-SSD	86.34	87.83	89.96	90.34	92.21
Proposed	**87.2**	**88.92**	**90.34**	**91.7**	**92.69**

Implementation. Following protocols commonly used in [5], [26], and [27], we train the proposed deep model on the Image VID and the DET. The same classes of the DET are selected for training. To output more useful candidate boxes, the IoU threshold of NMS is set to 0.7. Additionally, the SSD-based network is trained with an SGD optimizer, and the ST-LSTM-based network is trained with an ADAM optimizer [28]. The initial learning rate is 0.0001, the decay rate is 0.1, and the epoch number is 40.

B. Ablation study

Performance of various designs. We evaluate the effectiveness of multiple components in the proposed model, including the traditional SSD, the proposed SSD-based structure, the combination of the conventional LSTM, and the proposed model. As shown in Table 1, key components are tested on the ImageNet VID dataset. First, a 6.8% increase after the improved ST-LSTM is introduced, which is a significant improvement. This shows that attention maps effectively suppress useless information and temporal feature mining between frames, which is the main contribution in work.

C. Comparison with other state-of-the-art works

ImageNet VID benchmark. As shown in Table 2, the result comparison between the proposed method and other state-of-the-art methods has been summarized. Most methods are two-stage detectors based on the RPN, and few works achieve global-local temporal information mining using attentional-based components. Although the proposed model did not achieve the highest mAP, it is more concise than other models. For example, a tracking module is designed to help detect objects in the TSSD-OTA, making the model more complicated.

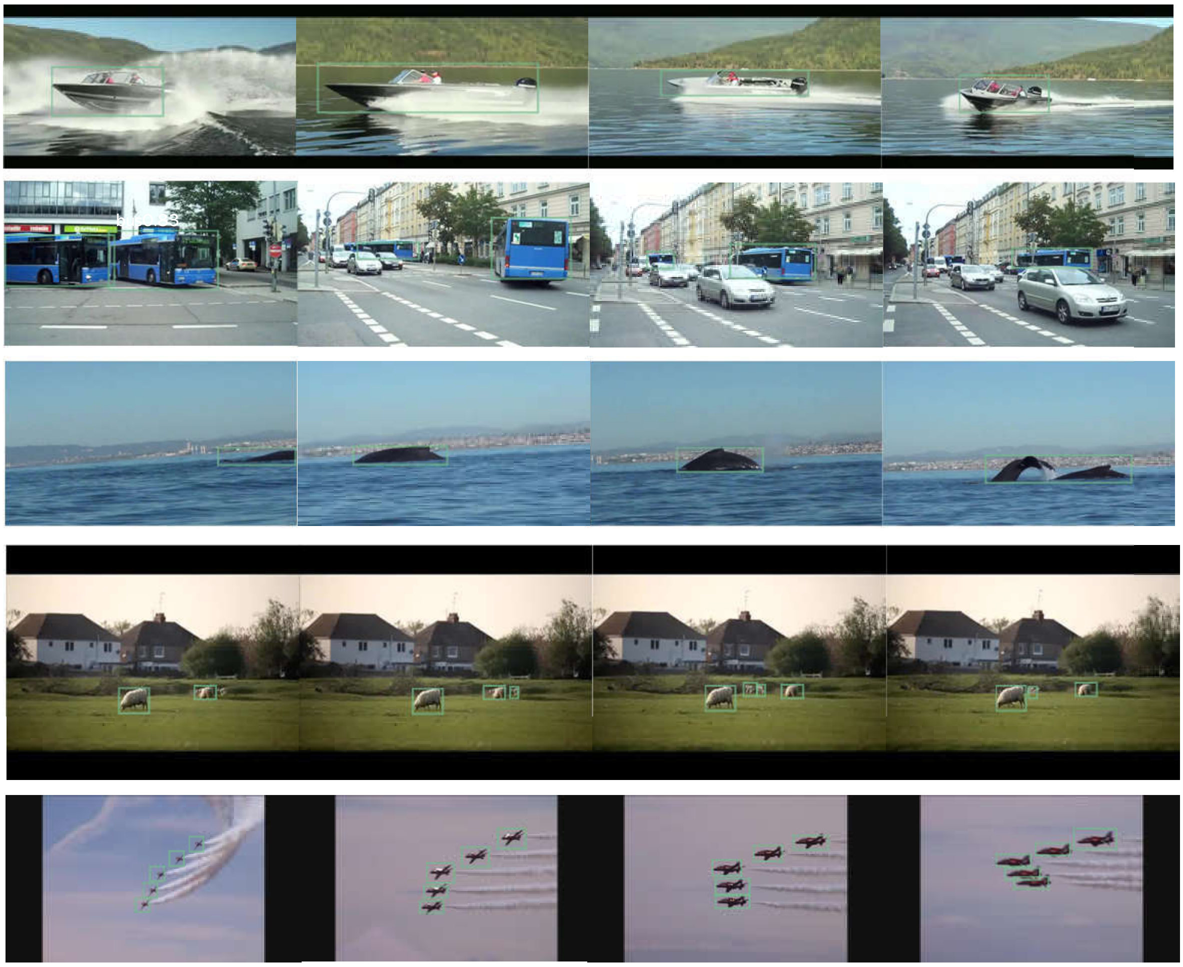
SRSSL dataset. The proposed method is evaluated under a common setting, and the comparison results are shown in Table 3. It can be seen that the proposed model achieves the best performance for the five sizes of hand detection and also accurately recognize eight common gestures. Referring to the FF-SSD [23], although it has achieved state-of-the-art detection accuracies, it is not good for small object detection and complex gesture recognition.

Some qualitative results are shown in Fig. 2. It can be seen that the proposed model can successfully detect multiclass objects in various challenging scenes, including complex backgrounds, occlusions, small pixel ratios, and blurring.

IV. CONCLUSIONS

In this paper, we propose a new one-stage temporal detector to achieve online object detection in videos. The proposed hierarchical structure can take full use of temporal information from inter-frame and intra-frame, which suppresses useless background information. The concise system achieves high-frame-rate object detection, which is a novel backbone for future studies. The proposed model achieves significant results on the ImageNet benchmark and the SRSSL dataset.

In future research, we plan to improve the detection frame rate and introduce a tracking module. Next, we will



focus on multiclass object classification and multisize object detection.

ACKNOWLEDGMENT

The authors would like to acknowledge the support from the AiBle project co-financed by the European Regional Development Fund, National Key R&D Program of China (Grant No. 2018YFB1304600), CAS Interdisciplinary Innovation Team (Grant No. JCTD-2018-11), LiaoNing Province Higher Education Innovative Talents Program Support Project (Grant No. LR2019058), and National Natural Science Foundation of China (grant No. 52075530 and 51575412).

REFERENCES

[1] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *arXiv preprint arXiv:1605.06409*, 2016.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[5] X. Chen, J. Yu, and Z. Wu, "Temporally identity-aware ssd with attentional lstm," *IEEE transactions on cybernetics*, vol. 50, no. 6, pp. 2674–2686, 2019.

[6] Y. Ji, H. Zhang, Z. Jie, L. Ma, and Q. J. Wu, "Casnet: a cross-attention siamese network for video salient object detection," *IEEE transactions on neural networks and learning systems*, 2020.

[7] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 337–10 346.

[8] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 817–825.

[9] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2349–2358.

[10] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 408–417.

[11] F. Xiao and Y. J. Lee, "Video object detection with an aligned spatial-temporal memory," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 485–501.

[12] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7210–7218.

[13] M. Shvets, W. Liu, and A. C. Berg, "Leveraging long-range temporal relationships between proposals for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9756–9764.

[14] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 879–888.

[15] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *arXiv preprint arXiv:1506.04214*, 2015.

[16] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Spatio-temporal closed-loop object detection," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1253–1263, 2017.

[17] H. Wu, Y. Chen, N. Wang, and Z.-X. Zhang, "Sequence level semantics aggregation for video object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9216–9224.

[18] M. Zhu and M. Liu, "Mobile video object detection with temporally-aware feature maps," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5686–5695.

[19] F. Xiao and Y. J. Lee, "Video object detection with an aligned spatial-temporal memory," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 494–510.

[20] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2018.

[21] X. Chen, J. Yu, and Z. Wu, "Temporally identity-aware ssd with attentional lstm," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2674–2686, 2020.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[23] Q. Gao, J. Liu, and Z. Ju, "Robust real-time hand detection and localization for space human–robot interaction based on deep learning," *Neurocomputing*, vol. 390, pp. 198–206, 2020.

[24] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9217–9225.

[25] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Relation distillation networks for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7023–7032.

[26] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang *et al.*, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2017.

[27] K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang, "Object detection in videos with tubelet proposal networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 727–735.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.