Adaptive Feature Aggregation for Video Object Detection

Yijun Qian Lijun Yu Wenhe Liu Guoliang Kang Alexander G. Hauptmann
Language Technologies Institute
Carnegie Mellon University
yijunqia@andrew.cmu.edu, lijun@cmu.edu, {wenhel, gkang}@andrew.cmu.edu alex@cs.cmu.edu

Abstract

Object detection, as a fundamental research topic of computer vision, is facing the challenges of video-related tasks. Objects in videos tend to be blurred, occluded, or out of focus more frequently. Existing works adopt feature aggregation and enhancement to design video-based object detectors. However, most of them do not consider the diversity of object movements and the quality of aggregated context features. Thus, they can not generate comparable results given blurred or crowded videos. In this paper, we propose an adaptive feature aggregation method for video object detection to deal with these problems. We introduce an adaptive quality-similarity weight, with a sparse and dense temporal aggregation policy, into our model. Compared with both image-based and video-based baselines on ImageNet and VIRAT datasets, our work consistently demonstrates better performance. Especially, our model improves the average precision of person detection in VIRAT from 85.93% *to* 87.21%*. Several demonstration videos of this work are available*[1].

1. Introduction

Object detection has grown into a fundamental field in the area of computer vision. It has been proved successful in providing detailed analysis of objects in images for various downstream tasks. With the emerging of video-related tasks, object detection is also developing from image to video. Compared with images, video frames bring along the following challenges for directly applying traditional image-based object detection models, as shown in Figure 2.

1. Motion blur caused by the fast-moving of objects.
2. Objects occluded by surroundings or other objects.
3. Objects out of focus due to camera movement.

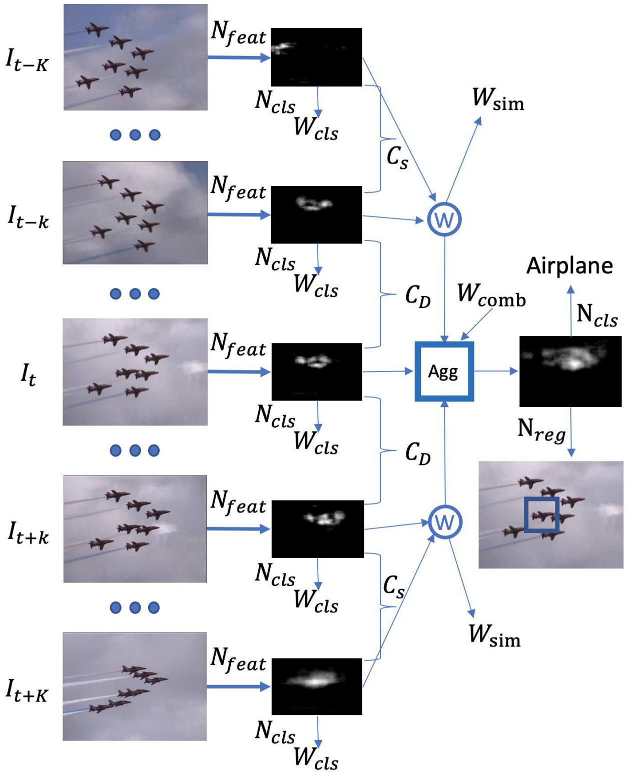
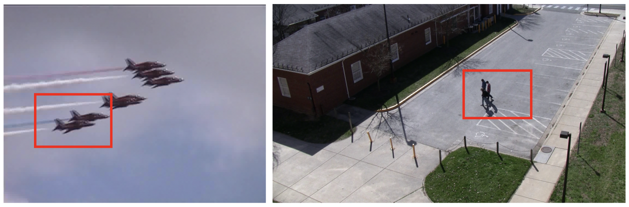


Figure 1. Model architecture for video object detection with adaptive feature aggregation, details in Section 3

Video object detection models utilize feature aggregation from context frames to overcome these challenges and get better performance. Nevertheless, no existing method has taken the diversity of object movements and the quality of context features into account, to the best of our knowledge. Almost all these methods use a fixed temporal window size for all objects and treat all context frames equally. However, in a complex scene with lots of objects, the optimal length of the temporal window for feature aggregation varies. For example, in a crowded scene where people are moving at various speeds, the window size should be large for a person temporarily occluded by the others. On the other hand, the

[1] https://drive.google.com/drive/folders/1U3o1m5tJ1PGW1PNk4N_oEsZosezuxtmc?usp=sharing



Motion Blur

Occlusion

Figure 2. Examples of motion blur and occlusion

model should only consider near frames for a fast-moving person. Moreover, if the object is blurred, occluded, or out of focus in a context frame, it should be assigned smaller weights in that it contains less useful information.

We state the major contributions of our work as follows:

1. We propose an adaptive feature aggregation method for video object detection, which combines weights of both similarity and quality.
2. We design a dense and sparse cache strategy for temporal feature aggregation, which further improve the VOD performance.
3. Our model outperforms the baselines on both ImageNet and VIRAT datasets. Especially, our model improves the average precision of person detection in VIRAT from 85.93% to 87.21%.

The rest of this work is organized as follows: In Section 2, we revisit the object detection models based on both image and video, and analyze their problems. In Section 3, we propose our method as improvements upon the baseline model. In Section 4, we provide experiment details and analysis to support our method. In Section 5, we conclude our work.

2. Background

Image based object detection (IOD) models such as Faster R-CNN[9] and R-FCN[3] have demonstrated convincing performance in video-related tasks such as multiple object tracking[10], event detection[2, 12] and danger recognition[13]. Given an image I as input, an IOD model usually uses a feature network N_{feat} to extract features as $f = N_{feat}(I)$. Based on the extracted features, it introduces a sub-network N_{det} for detection, which generates a label y and a bounding box b. This classical structure has advanced the state of the arts in various challenges. However, for object detection in continuous videos, objects like airplanes may get blurred due to their high-speed movement. Meanwhile, in crowded scenes, objects might get occluded by other moving objects frequently. Thus, IOD models that only use a single frame usually generate unstable results or fails on these videos.

Directly applying IOD models frame by frame would suffer from blur, occlusion, and out of focus problems, which are common in videos. To deal with the continuous scenario, many approaches have been developed for video object detection (VOD) with feature propagation and enhancement modules. Zhu *et al*. proposed a flow-guided feature aggregation (FGFA) method[15] and multi-frame end-to-end learning of features with cross-frame motion[14]. Xiao *et al*. introduced an aligned spatial-temporal memory network to model temporal appearance and motion dynamics[11]. Hetang *et al*. also implemented an impression network to balance the efficiency and accuracy of detection[6]. Different from previous works which based on optical flow for alignments, Bertasius *et al*. implements Deformable Convolution[1] and Xiao *et al*. resorts to MatchTrans [11]. Generally, these models generate an alignment matrix $A(I_t, I_{t'})$ to combine the context frames into a target frame. Given the feature extracted from target frame $f_t = N_{feat}(I_t)$ and a context frame $f_{t'} = N_{feat}(I_{t'})$ with alignment $A(I_t, I_{t'})$, a warp module W propagates $f_{t'}$ to $f_{t'}^t = W(f_{t'}, A(I_t, I_{t'}))$. Then the enhancement module aggregates these warped features and generate final results.

Although these VOD models noticed the importance of using contextual information, they did not design specific strategies for different objects and their diverse requirements of temporal length for reference. Meanwhile, during aggregation, they do not take the quality of contextual features into consideration and may suffer from low-quality contexts.

3. Methodology

We intend to integrate the baseline FGFA model with an adaptive feature aggregation method to deal with the problems aforementioned.

3.1. Adaptive Quality-Similarity Weight

In the aggregation module, FGFA only focuses on the similarity between warped features and features extracted from the target frame. However, if the object on the target frame is blurred or occluded, similar blurred or occluded frames will be assigned higher weights due to the similarity weight. On the other hand, the clear or complete frame will get lower weights. To solve this, we propose a combination weight W_{comb} which contains both similarity weight W_{sim} and classification weight W_{cls}, as is shown in Equation 1-3. W_{sim} is directly calculated as the cosine distance of the extracted features. W_{cls} is a pixel-wise weight where $W_{cls}^{(i,j)}$ equals to the maximum probability of the Region of Interests (ROIs) which contain pixel (i, j), divided by the number of these ROIs. And the probabilities of ROIs are generated by sub-network N_{cls}.

$$
\begin{aligned}
W_{sim}^{t^{'}\to t} &= \exp\left(\frac{f_t \cdot f_{t'}^t}{|f_t| \times |f_{t'}^t|}\right) &(1)\\
W_{cls}^{t^{'}\to t} &= \frac{\max_{ROIs} N_{cls}(f_t)}{N_{ROIs}} &(2)\\
W_{comb}^{t^{'}\to t} &= softmax(W_{sim}^{t^{'}\to t} \times W_{cls}^{t^{'}\to t}) &(3)
\end{aligned}
$$

3.2. Dense and Sparse Temporal Aggregation

Within a complex scene such as crowded people, objects need different lengths of contextual frames for reference. For a small fast-moving object, its appearance may change drastically, and warping from a far frame may cause errors. Therefore, we just need to refer to close dense frames. On the other hand, for continuously occluded or blurred objects, such as people talking behind a vehicle, we need to resort to further frames for visibility. Based on this, we design a dense and sparse strategy for temporal feature aggregation. In this method, two feature caches are stored with different frame gaps and temporal window sizes. For the dense one C_D, we store close contextual frames one by one. For the sparse one C_S, we store further contextual frames with a frame gap g. Given cached features and the adaptive weight matrix, we generate the aggregated feature of the target frame as:

$$f_{agg}^t = \sum_{f_i \in C_D} W_{comb}^{t^i \to t} f_i^t + \sum_{f_j \in C_S} W_{comb}^{t^i \to t} f_j^t \qquad (4)$$

3.3. Model Architecture

With the baseline and sub-modules illustrated above, the structure of our model is shown in Figure 1. We store features extracted by N_{feat} in caches C_S and C_D and update the classification weights W_{cls}. Then we warp the contextual features to the target frame and calculate the similarity weights W_{sim}. After that, we calculate combination weights W_{comb} with updated W_{sim} and W_{cls}, and generate f_{agg}^t through aggregation. Finally, N_{cls} gives out the classification probability and N_{reg} gives out the bounding box locations based on the aggregated feature f_{agg}^t.

4. Experiments

4.1. Datasets

We conduct experiments on a large image dataset ImageNet2015[4] and a challenging surveillance dataset VIRAT[8].

The ImageNet2015 is a large-scale dataset for object detection and classification. It provides annotations of both image dataset DET and video dataset VID. There are 200 object categories in DET and 30 categories in VID, which is a subset of DET.

The VIRAT dataset is collected in natural scenes showing people performing normal actions in standard contexts, with uncontrolled, cluttered backgrounds. Different from ImageNet2015, there are a large number of objects with different speeds and sizes within a single frame with frequent occlusions.

4.2. Implementation Details

For the IOD baseline, we followed FGFA to adopt an R-FCN with ResNet 101[5] pre-trained on ImageNet as the backbone network.

For training, the entire network is fully differentiable for an end to end optimization. Since we use the softmax function to normalize the adaptive weights for each pixel, we can repeatedly use N_{cls} with frozen weights when calculating w_{cls}. Due to the limit of GPU memory, we set the size of both C_D and C_S as 2. As a result, we will randomly aggregate two features from each cache in each iteration.

For inference, we have more GPU memory for larger feature caches, which contain $2K$ features each. Given an input video, we first initialize the feature caches with K copies of f_1. After that, we iteratively load images and update the caches and weights matrix. Once the size of each feature cache reaches $2K$, we start to perform inference based on the feature extracted from the target frame, feature caches, and weights matrix. When loading the last frame I_N, we just make the inference of frame I_{N-k}. Similar to the initialization, we update feature caches and weights matrix continuously with f_N until the inference for all frames is finished.

Given that the 30 object categories in VID are a subset of 200 categories in DET, we can use both VID and these 30 object categories in the DET set for training. To be specific, in each epoch, we firstly train N_{feat}, N_{cls} and N_{reg} with DET data. Then we train the whole model on VID where the N_{feat}, N_{reg} and N_{cls} are initialized with weights learned on DET. For ImageNet2015, we resize the shorter size of all images to 600 pixels. The learning rate is set as 2.5^{-4} and deteriorate to 2.5^{-5} after 6 epochs. For VIRAT, we resize the shorter size of all images to 860 pixels. The learning rate is set as 1^{-3} and deteriorate to 5^{-4} after 5 epochs. Since many object categories of VIRAT overlap with the COCO[7] dataset, we pre-trained N_{feat}, N_{cls} and N_{reg} on COCO. Meanwhile, given that VIRAT contains much more small objects, we changed the anchor size from $[8, 16, 32]$ to $[2, 4, 8, 16, 32]$.

4.3. Performance Comparison

For evaluation, we compare our model with the IOD baseline and FGFA on ImageNet. Results in Table 1 demonstrate that our model outperforms the two baselines. Compared with FGFA, the mean average precision at the threshold of 0.5 (mAP@0.5) is improved from 77.1% to 77.9%.

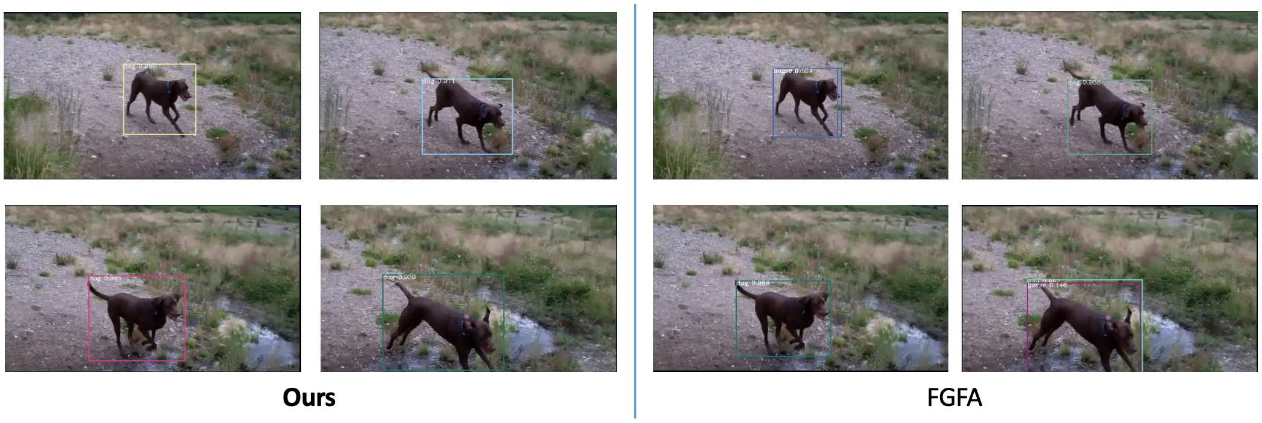


Figure 3. The performance of our model and FGFA on an example video clip. FGFA sometimes misclassifies the dog into a horse whereas ours generates the correct results.

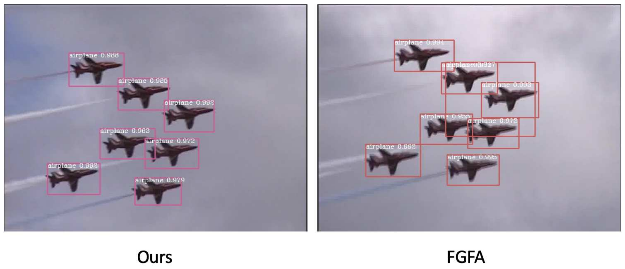


Figure 4. The performance of our model and FGFA on an example video clip. Ours generates more accurate bounding box than FGFA.

The improvements of adaptive weights and feature aggregation in our model achieve overall performance promotion. Especially, as is shown in Figure 3 and 4, our model can generate more reliable classification results and more precise bounding boxes.

Table 1. Comparison with Baseline Works on ImageNet

Model	mAP@0.5
R-FCN	74.1%
FGFA	77.1%
Ours	**77.9%**

According to the results in Table 2, our model also outperforms the other two baselines on VIRAT. Since VIRAT only contains human actions and human-vehicle interactions, we calculate the average precision (AP) of person and vehicle separately in addition to mAP@0.5. The results in Table 2 shows that our model improvements the AP of both person and vehicle. The AP of person is significantly improved from 85.93% to 87.21%. As is shown in Figure 5, our model has considerable improvements when objects get occluded.

Table 2. Comparison with Baseline Works on VIRAT

Model	mAP@0.5	AP Person	AP Vehicle
R-FCN	70.21%	0.8527	0.9092
FGFA	73.79%	0.8593	0.9111
Ours	**74.27%**	**0.8721**	**0.9121**

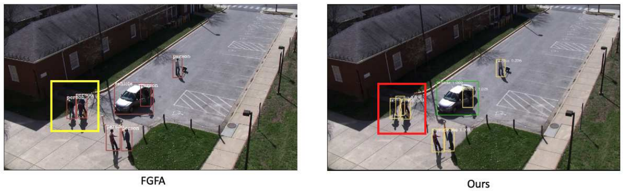


Figure 5. Our model successfully detects all three overlapped people in the red box whereas FGFA only detects two.

5. Conclusion

In this paper, we propose an accurate video object detection model with adaptive feature aggregation. Compared with previous works, it calculates an adaptive quality-similarity weight of context frames and integrates a dense and sparse temporal aggregation policy. According to the results on ImageNet and VIRAT, our model demonstrates better performance in comparison with the baseline IOD and FGFA models on both datasets. Especially, our model improves the average precision of person detection in VIRAT from 85.93% to 87.21%. It shows that our improvements can promote overall performance, with a boost in specific categories. Since our model does not have special requirements for the structure of N_{feat} and N_{det}, it can be easily generalized to majority image-based object detection models. Several aspects are left for future explorations. For example, more precise alignment methods and long-term feature aggregation scheme.

6. Acknowledgment

This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation/herein.Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements,either expressed or implied, of IARPA, DOI/IBC, or the U.S.Government.

References

[1] G. Bertasius, L. Torresani, and J. Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018.

[2] X. Chang, W. Liu, P.-Y. Huang, C. Li, F. Zhu, M. Han, M. Li, M. Ma, S. Hu, G. Kang, et al. Mmvg-inf-etrol@ trecvid 2019: Activities in extended video.

[3] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] C. Hetang, H. Qin, S. Liu, and J. Yan. Impression network for video object detection. *arXiv preprint arXiv:1712.05896*, 2017.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[8] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[10] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

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

[12] L. Yu, P. Chen, W. Liu, G. Kang, and A. G. Hauptmann. Training-free monocular 3d event detection system for traffic surveillance. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019.

[13] L. Yu, D. Zhang, X. Chen, and A. Hauptmann. Traffic danger recognition with surveillance cameras without training data. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[14] 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*, pages 7210–7218, 2018.

[15] 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*, pages 408–417, 2017.