VidVRD 2021: The Third Grand Challenge on Video Relation Detection

Wei Ji

Sea-NExT Joint Lab
National University of Singapore
jiwei@nus.edu.sg

Yicong Li

National University of Singapore
liyicong@u.nus.edu

Meng Wei

National University of Singapore
t0923065@u.nus.edu

Xindi Shang

National University of Singapore
shangxin@comp.nus.edu.sg

Junbin Xiao

National University of Singapore
junbin@comp.nus.edu.sg

Tongwei Ren

Nanjing University
rentw@nju.edu.cn

Tat-Seng Chua

National University of Singapore
chuats@comp.nus.edu.sg

ABSTRACT

ACM Multimedia 2021 Video Relation Understanding Challenge is the third grand challenge which aims at exploring the relationship of subjects and objects appearing in videos for fine-grained and high-level video understanding. Given a video, the video relation detection model should output a series of relation triplet <subject, predicate, object> and the corresponding trajectories of subject and object. The goal of this task is to promote research on developing video semantic understanding model, so as to perform complex inferences and mining of visual knowledge in videos. In this paper, we make a comprehensive and detailed introduction of this task, conclude the proposed algorithms in the last few years, and propose future direction for research in this task.

CCS CONCEPTS

• Computing methodologies → Activity recognition and understanding.

KEYWORDS

Video Relation Detection, Object Trajectory, Relation Prediction

ACM Reference Format:

Wei Ji, Yicong Li, Meng Wei, Xindi Shang, Junbin Xiao, Tongwei Ren, and Tat-Seng Chua. 2021. VidVRD 2021: The Third Grand Challenge on Video Relation Detection. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21)*, Oct. 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3474085.3479232>

1 INTRODUCTION

With the impressive development of deep learning, there has been great progress in multimedia analysis system, such as the object

Figure 1: Example of video relation detection task. Given a video, trajectories of subjects (in green bounding box) and objects (in red bounding box) shall be extracted first, and then the relationship (e.g. chase, bite) of subjects and objects in different sequences need to be produced by inference. Then the output of the video relation detection task should be in the form of (<subject,predicate,object>) and trajectories of the subjects and objects. Adapted from the original paper [19].

Figure 2: Examples of visual relation in images and videos. The top figure shows the visual relations in image, which includes the relation triplets and object localization (with bounding boxes). The bottom figure presents video relation detection. Compared with image-level visual relation, video relation detection covers more abundant relations with temporal information, and the object localization of objects are in the form of trajectories within consecutive frames [19].

This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '21, October 20–24, 2021, Virtual Event, China.

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8651-7/21/10.

<https://doi.org/10.1145/3474085.3479232>

tracking [12], semantic segmentation [9][8][10], etc. However, high-level semantic understanding in videos is still a challenging task.

Figure 3: The pipeline of video relation detection [19]. The whole framework consists of three parts: object tracklet proposal, relation prediction, and greedy relational association. In the object tracklet proposal, the given video is first divided into several segments with partial overlapping. Then, the trajectories of objects in each segment are extracted as the object localization information with bounding box, along with the category information of subjects and objects. In the relation prediction part, the relations of subjects and objects are predicted by the proposed relation model. Finally, the relation of the whole video are merged with greedy relational association algorithm.

Fine-grained visual content understanding and extraction is crucial but lack of exploration. To bridge the gap between visual and language, and enhance the interpretability of multimedia analysis system, visual relation detection (VRD) is proposed for further research.

Early works in visual relation detection mainly focus on static images (ImgVRD) [13][3] and achieve satisfactory results. However, videos contain more complex relations with temporal information, as shown in Figure 2. Since the relations in video are more complex with temporal information, video relation detection is more challenging than the image counterpart. The relation triplets <subject, predicate, object> represent the salient relationship of subjects and objects in consecutive frames as shown in Figure 1. And the research on video relation detection will benefit a lot of downstream tasks, such as video classification [30], video recommendation [23][24], video retrieval[28][26], video question answering [11] and many other visual-and-language tasks [4].

In this paper, we present a comprehensive and detailed introduction of VRU task, summarize the current proposed solutions, and propose the future direction of research for this task.

2 APPROACH

Most proposed solutions are comprised of three parts: object tracklet proposal, relation detection, and greedy relational association, as shown in Figure 3.

2.1 Object Tracklet Proposal

Object tracklet proposal plays a fundamental role in the VRD task, which heavily contribute to the performance of relation prediction.

In [22], Sun et.al. train the FGFA model [33] with ResNet-101 network [7] as the backbone. After filtering out the detection result with low confidence, they propose some heuristic rules and apply Seq-NMS [6] to generate the short-term preliminary trajectories.

After another filtering, trajectories with confidence higher than the threshold are reserved for relation prediction.

Besides, Chen et.al.[14] first employ an object detector in video frames, then track frame-level detection results across the whole video. In detail, they use a pretrained Faster-RCNN [16] with ResNet-101 [7] as image detector. For object tracking, they choose deep sort [25] as their tracker so as to integrate visual features as matching descriptors. The visual feature in deep sort are object detection features obtained from RoI Pooling. Afterwards, they adopt a sliding-window approach to generate object tracklet proposals. To filter out incompatible proposal pairs, they construct spatial and temporal graph, and use Graph Convolutional Networks to embed the spatial and temporal contextual information for each tracklet proposal.

To obtain better performance, Xie et.al. [27] first perform object detection for each video frame by Cascade R-CNN [1] with ResNeSt101 [29] as the backbone and use a Dynamic Programming Algorithm extended from seq-NMS [6] to generate the trajectories for each object.

2.2 Relation Prediction

After obtaining the object tracklet proposal, several methods follow the bottom-up strategy as [19], which consists of three steps: relation candidate breaking, predicate recognition on segments and relation segment association.

For each object trajectory pair, Sun et.al. [22] break the co-occurrent part into segments with fixed duration as relation candidate segments. They then combine the relative location feature, motion feature, and language context feature to make final relation prediction, connected with fully-connected neural networks and softmax layer.

In [14], Chen et.al. extract the subject/object feature and relative motion feature as the relation feature, followed by fully-connected layers to predict the relation result of object tracklet proposals.

Finally, they utilize multi-label classification and binary cross entropy loss function to calculate the loss between the prediction and groundtruth.

Different from the methods above, Xie et.al. [27] combine the motion feature, visual feature, language feature and location mask feature for each trajectory pair to predict the relation between them. The utilization of multi-modal feature helps a lot to increase the accuracy of relation prediction.

2.3 Greedy Relational Association

After obtaining the relation prediction results for all pairs of object tracklet proposals at the video segment level, Shang et.al. [19] propose a relational association algorithm to merge the relations detected in segments. The proposed greedy algorithm for relational association repeatedly merges two most confident visual relation instances that overlap in two successive segments. The greedy strategy can help to generate visual relation instances with longer sequence, so as to make the subject and object in each visual relation to be temporally localized more accurately.

By merging all possible visual relation instances, Shang et.al. [19][18] rank them according to their confidence scores and output the most confident ones as the visual relation detection results for the video.

3 DATASET

As shown in Table 1, there are two datasets mainly used to evaluate the performance of video relation detection: ImageNet-VidVRD [19] and VidOR [17] dataset. ImageNet-VidVRD dataset is collected from ILVSRC2016-VID dataset, which contains 800 videos for training and 200 videos for testing. Each selected video contains clear visual relations. This dataset provides well-labeled relations of objects and corresponding accurate trajectories, which covers 35 categories of object and 132 types of relation.

Another video relation detection dataset is VidOR [17], which comprise 98.6 hours of user-generated videos collected from YFCC100M dataset. The annotation of VidOR is in the same format as ImageNet-VidVRD. The average length of videos is about 35 seconds. The training, validation and testing set of VidOR has 7,000, 835, and 2,165 videos, respectively. Objects in 80 categories are annotated with precise trajectories with bounding boxes and the corresponding relations in 50 categories. The training and validation data are publicly available, which can be downloaded from the Project page¹, and the testing set is open for evaluation of annual grand challenge.

4 EVALUATION METRIC

We adopt Average Precision (AP) to evaluate the detection performance per video and finally calculate the mean AP (mAP) over all testing videos as the ranking score. In particular, we calculate the AP similar to that in the Pascal VOC challenge. To match a predicted relation instance ($\langle s, p, o \rangle^p, (\mathcal{T}_s^p, \mathcal{T}_o^p)$) to a ground truth ($\langle s, p, o \rangle^g, (\mathcal{T}_s^g, \mathcal{T}_o^g)$), we require:

- (1) their relation triplets to be exactly the same, i.e. $\langle s, p, o \rangle^p = \langle s, p, o \rangle^g$;

- (2) their bounding-box trajectories overlap s.t. $\text{vIoU}(\mathcal{T}_s^p, \mathcal{T}_s^g) \geq 0.5$ and $\text{vIoU}(\mathcal{T}_o^p, \mathcal{T}_o^g) \geq 0.5$;
- (3) the minimum overlap of the subject trajectory pair and the object trajectory pair $\text{ov}_{pg} = \min(\text{vIoU}(\mathcal{T}_s^p, \mathcal{T}_s^g), \text{vIoU}(\mathcal{T}_o^p, \mathcal{T}_o^g))$ is the maximum among those paired with the other unmatched ground truths $\mathcal{G}$, i.e. $\text{ov}_{pg} \geq \text{ov}_{pg'} (g' \in \mathcal{G})$.

The term vIoU refers to the voluminal Intersection over Union.

While calculating the score, we only consider the top-200 predictions for each video.

5 PERFORMANCE COMPARISON

We compare current solutions proposed for video relation detection task: TMF [27], RELAbuilder [32], MAGUS.Gamma [22], VRD-STGC [14], and so on. Since the test set is only open for participates of VRU grand challenge, the testing result of some methods are not available.

As shown in Table 2 and Table 3, TMF [27] achieves the best performance. This is because their video object detection method actually outperform the other methods by a large margin. The accurate trajectories also benefits the prediction of relations. More details can be found in their original paper [27].

6 FUTURE DIRECTION

Although there has been great progress in the research of video relation, the performance of state-of-the-art VRD methods is still far from satisfactory. Hence, we propose the several directions for future exploration.

Stronger Baseline for Object Tracklet Proposal. The current solutions mainly consist of three parts: object tracklet proposal, relation prediction, and greedy relational association.

If we don't consider the accuracy of trajectories and only take the accuracy of relation into consideration, which is the relation tagging task, the state-of-the-art method can achieve more than 70% in accuracy [27]. However, the accuracy of relation detection (with the evaluation of trajectories accuracy) is only 11.74%, which means the quality of trajectory generation is the current bottleneck of the video relation task. Also, as confirmed by previous works [27] [14], the performance of relation prediction will improve correspondingly if given better trajectories. Recently, a serious of object detectors and trackers has been proposed, such as DETR [2], TrackFormer [15]. Hence, a stronger object tracking baseline for video relation detection shall be developed.

Complement with Multi-modal Data. The performance of current solutions heavily rely on the annotations. That is the reason why a larger video dataset VidOR is proposed after the release of ImageNet-VRD. However, the cost of annotating a video relation dataset is very expensive and time-consuming, due to the accurate dense tracklet labels in the video. However, there also exist data with relation information in other modal (such as text, audio, etc), utilizing these data can relieve the dependency of labels in video data. For example, we can extract the relation information of subject and object from large paragraphs of text description, which is easier to obtain than the dense video labels.

Deal with labels in Long-tailed Distribution. As shown in Figure 4, the relation of whole video is in long-tailed distribution.

¹<https://xdshang.github.io/docs/vidor.html#downloads>

Figure 4: The annotations of relations in video relation detection result in the long-tailed distribution [14]. The unbalanced labeling brings difficulty in training robust models.

Table 1: Statistics of two datasets used for video relation benchmark. #Obj-Category and #Obj-Instance denote the total number of object categories and instances in the datasets, respectively [20].

Dataset	Hours	Train	Val	Test	Obj-Category	Obj-Instance	Predicate Category	Relation Instance
ImageNet-VidVRD [19]	3.0	800	-	200	35	3,017	132	4,835
VidOR [17]	98.6	7,000	835	2,165	80	49,258	50	378,546

Table 2: Comparison with state-of-the-art methods on VidOR validation set (%).

Method	tagging P@1	tagging P@5	R@50	R@100	mAP
RELAbuilder[32]	33.05	35.27	1.58	1.85	1.47
OTD+GSTEG[22]	51.20	37.26	6.40	8.43	5.58
OTD+CAI[22]	48.31	38.49	6.19	8.16	5.65
MAGUS.Gamma[22]	51.20	40.73	6.89	8.83	6.56
VRD-STGC[14]	48.92	36.78	8.21	9.90	6.85
TMF[27]	67.43	-	9.12	-	9.93

Table 3: Comparison with state-of-the-art methods on VidOR test set (%).

Method	mAP	R@50	R@100	tagging P@1	tagging P@5	tagging P@10
DeepBlueAI	0.24	-	-	-	-	-
RELAbuilder[32]	0.55	-	-	-	23.60	-
GKBU	3.28	-	-	-	-	-
MHA [21]	5.99	-	-	-	-	-
MAGUS.Gamma [22]	6.31	-	-	-	42.10	-
ETRI_DGRC	6.65	-	-	-	-	-
Planck	6.69	7.13	9.34	54.22	40.96	32.51
Item-XiXi [5]	9.47	8.56	10.44	63.46	54.07	41.95
TMF [27]	11.74	10.02	12.69	71.36	56.30	44.59

Relations (such as throw, knock) rarely appear in the video annotations. How to deal with the long-tailed problem is a promising direction. Besides, relation detection in the setting of incremental learning and few-shot learning [31] can also be explored.

7 CONCLUSION

It has been three years since we host the Video Relation Understanding Challenge in ACM Multimedia Conference. This challenge aims at promoting research on developing video semantic understanding model, so as to perform complex inferences and mining

of knowledge in videos. In this paper, we have given a comprehensive and detailed introduction of this task, concluded the proposed algorithms in the last few years, and proposed future direction of research in this task. Also, we will collaborate with ACM Multimedia and other top conference community to organize the following VRU grand challenges.

ACKNOWLEDGEMENT

This research is funded by Sea-NEXt Joint Lab, Singapore.

REFERENCES

- [1] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.
- [3] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. 2019. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4613–4623.
- [4] Hehe Fan, Linchao Zhu, Yi Yang, and Fei Wu. 2020. Recurrent attention network with reinforced generator for visual dialog. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 3 (2020), 1–16.
- [5] Kaifeng Gao, Long Chen, Yifeng Huang, and Jun Xiao. 2021. Video Relation Detection via Tracklet based Visual Transformer. In *ACM MM*.
- [6] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465* (2016).
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [8] Wei Ji, Xi Li, Lina Wei, Fei Wu, and Yueting Zhuang. 2020. Context-aware graph label propagation network for saliency detection. *IEEE Transactions on Image Processing* 29 (2020), 8177–8186.
- [9] Wei Ji, Xi Li, Fei Wu, Zhijie Pan, and Yueting Zhuang. 2019. Human-centric clothing segmentation via deformable semantic locality-preserving network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2019), 4837–4848.
- [10] Wei Ji, Xi Li, Yueting Zhuang, Omar El Farouk Bourahla, Yixin Ji, Shihao Li, and Jiabao Cui. 2018. Semantic Locality-Aware Deformable Network for Clothing Segmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. 764–770.
- [11] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. 2019. Multi-interaction network with object relation for video question answering. In *Proceedings of the 27th ACM international conference on multimedia*. 1193–1201.
- [12] Xi Li, Liming Zhao, Wei Ji, Yiming Wu, Fei Wu, Ming-Hsuan Yang, Dacheng Tao, and Ian Reid. 2018. Multi-task structure-aware context modeling for robust keypoint-based object tracking. *IEEE transactions on pattern analysis and machine intelligence* 41, 4 (2018), 915–927.
- [13] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. 2019. Natural language guided visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
- [14] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. 2020. Localize, Assemble, and Predicate: Contextual Object Proposal Embedding for Visual Relation Detection. In *CVPR*.
- [15] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2021. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702* (2021).
- [16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.
- [17] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 279–287.
- [18] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2021. Video Visual Relation Detection via Iterative Inference. In *ACM MM*.
- [19] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *Proceedings of the 25th ACM international conference on Multimedia*. 1300–1308.
- [20] Xindi Shang, Junbin Xiao, Donglin Di, and Tat-Seng Chua. 2019. Relation understanding in videos: A grand challenge overview. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2652–2656.
- [21] Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. 2020. Video Relation Detection via Multiple Hypothesis Association. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3127–3135.
- [22] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. 2019. Video visual relation detection via multi-modal feature fusion. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2657–2661.
- [23] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* 29 (2019), 1–14.
- [24] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.
- [25] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.
- [26] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2986–2994.
- [27] Wentao Xie, Guanghui Ren, and Si Liu. 2020. Video relation detection with Trajectory-aware multi-modal features. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4590–4594.
- [28] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. *arXiv preprint arXiv:2106.01534* (2021).
- [29] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. 2020. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955* (2020).
- [30] Hanbin Zhao, Xin Qin, Shihao Su, Yongjian Fu, Zibo Lin, and Xi Li. 2021. When Video Classification Meets Incremental Classes. *arXiv preprint arXiv:2106.15827* (2021).
- [31] Hanbin Zhao, Hui Wang, Yongjian Fu, Fei Wu, and Xi Li. 2021. Memory Efficient Class-Incremental Learning for Image Classification. *IEEE Transactions on Neural Networks and Learning Systems* (2021).
- [32] Sipeng Zheng, Xiangyu Chen, Shizhe Chen, and Qin Jin. 2019. Relation understanding in videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2662–2666.
- [33] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 408–417.