Performance Improvement Method of the Video Visual Relation Detection with Multi-modal Feature Fusion

Kwang-Ju Kim
Electronics and Telecommunications Research Institute
1 Techno Sunhwan-ro 10-gil, Yuga-eup, Dalseong-gun
Daegu, Korea 42994
kwangju@etri.re.kr

Pyong-Kun Kim
Electronics and Telecommunications Research Institute
1 Techno Sunhwan-ro 10-gil, Yuga-eup, Dalseong-gun
Daegu, Korea 42994
iros@etri.re.kr

Kil-Taek Lim
Electronics and Telecommunications Research Institute
1 Techno Sunhwan-ro 10-gil, Yuga-eup, Dalseong-gun
Daegu, Korea 42994
ktl@etri.re.kr

Jong Taek Lee
Electronics and Telecommunications Research Institute
1 Techno Sunhwan-ro 10-gil, Yuga-eup, Dalseong-gun
Daegu, Korea 42994
jongtaeklee@etri.re.kr

***Abstract*—Video visual relation detection is a novel research problem that aims to detect instances of visual relations of interest in a video. In this paper, we propose a performance improvement method of the video visual relation detection with multi-modal feature fusion. First, we introduce a spatial feature extraction method that is designed to include the relative positions of objects itself and between objects in the image. Next, we suggest a relationship classifier that is designed to accommodate the complexity of the input features. Our proposed method achieves 6.65 mAP, and ranked the 2nd place in the visual relation detection task of Video Relation Understanding Challenge (VRU), the ACM Multimedia 2020.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In recent years, deep learning technologies have achieved great success in computer vision tasks, such as object classification, detection, attribute detection, and segmentation [1]–[5]. These computer vision researches have improved the performance in various tasks of image understanding. However, high-level image understanding tasks such as image captioning, scene graph, visual question answering, image retrieval, and other related works remain open-challenging tasks [6]–[9]. As a mid-level learning task, visual relationship detection (VRD) can provide rich information for high-level image understanding tasks. A VRD is generally defined as a pair of objects localized by bounding-boxes together with a predicate to connect them. It aims to construct a holistic representation by identifying triplets in the form (subject, predicate, object). Comparing with VRD on static image, video visual relation detection (VidVRD) is much more practical and challenging than VRD. Firstly, dynamic interactions between objects can only be observed in videos. Secondly, there is a high variability of interactions between two specific objects in a video which causes another challenging problem. Therefore, if VRD methods used in the static image are applied to video, it is difficult to achieve high performance.

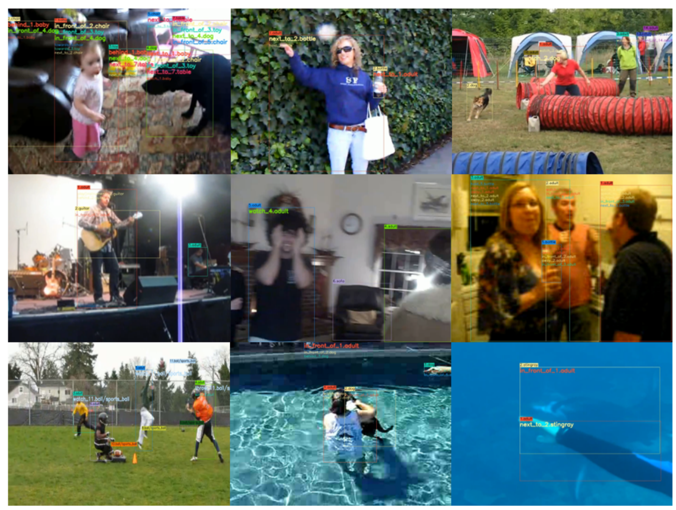

Fig. 1. Several examples of the VidOR dataset. Each object is spatio-temporally annotated, and the relation instances between each pair of objects are annotated in the videos.

To tackle these problems, several researches have been recently proposed. Their natural VidVRD's method is to generate features of dynamic and time-varying relationships between entities. On the other side, new challenges such as The ACM Multimedia 2019 VRU Challenge with Video Object Relation (VidOR) dataset [10], is designed to encourage this research. Figure 1 shows several examples of VidOR dataset, which contain 80 categories of objects annotated with a bounding-box trajectory to indicate their spatio-temporal location in the videos and 50 categories of relation predicates annotated

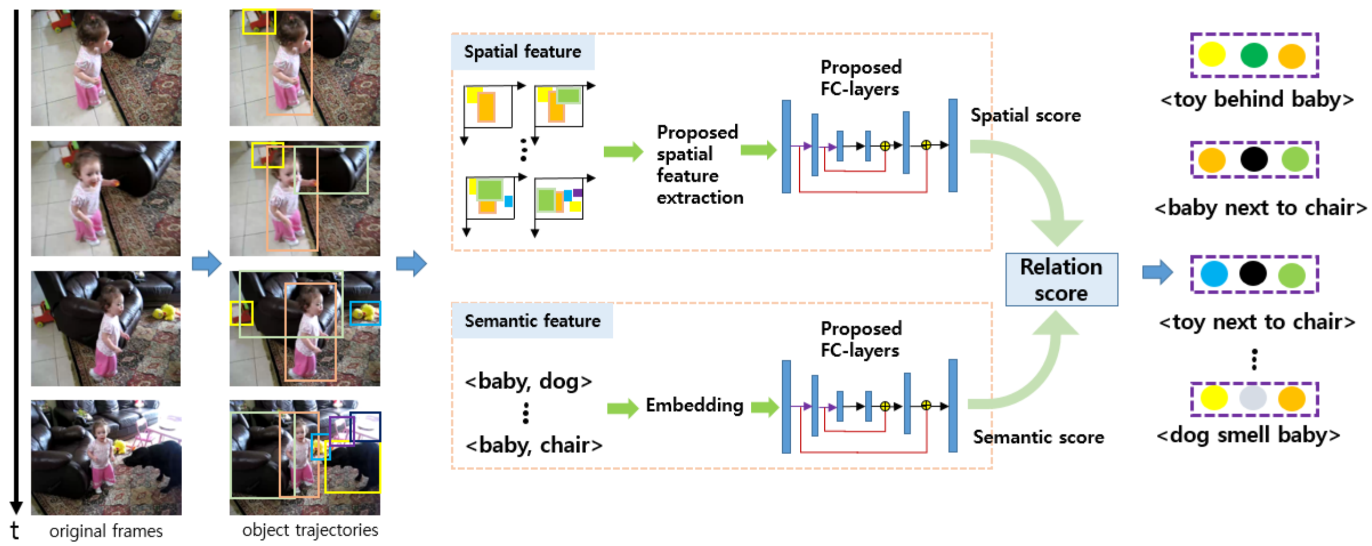

Fig. 2. The overview of our proposed model. The circles with different colors represent different predicates in the relation prediction results.

among all pairs of annotated objects with starting and ending frame index. The ACM Multimedia 2019 VRU Challenge winner proposed a multi-modal feature fusion method for VidVRD [11]. However, the winner's method still leaves room for performance improvements by modifying multi-modal feature extraction and relations prediction classification. In this paper, we propose a performance improvement method of video visual relation detection via multi-modal feature fusion.

II. Related Work

In recent years, many previous works have been studied the problem of the visual relationship prediction. We present the related work of video visual relation detection (VidVRD) as well as video object detection (VID).

A. Video Object Detection

VID is the task of detecting objects from a video as opposed to images. When image-based object detection methods applied to the video data, they can cause more miss detections because the appearance of objects becomes often blurred or even occluded in frames. After introducing the ImageNet video object detection challenge (ImageNet VID) [12], many object detection research efforts have been extended to video object detection. Many works utilized the idea of feature aggregation to enhance per-frame features by aggregating nearby frames' features. Specifically, Flow-Guided Feature Aggregation (FGFA) [13] utilizes an optical flow network from FlowNet [14], [15] for estimating the pixel-level motions on feature maps of adjacent frames for feature aggregation. Another solution to video object detection is to explore mapping strategies to link the static image detection results of the same object identity into a bounding-box trajectory. Seq-NMS [16] proposes a post-processing heuristic method consisting of three steps: sequence selection, re-scoring, and suppression. Through this method, the overall score was improved by correcting the score of weaker detection. Detect and Track (D&T) [17] generates a tracking formulation given two (or more) frames as input into R-FCN [18] to perform object detection and across-frame track regression. There is also proposed a method [19] to calibrate object feature at the box level to improve video object detection with an extended version of FGFA.

B. Visual Relation Detection

VRD aims to identify groups of objects and their relationships in an images in the form of (subject, predicate, object). Specifically, this task is to detect all objects presented in the image and predict all possible visual relationships between two of the detected objects. In the past few years, several approaches have been proposed to recognize the relationship from the static images. These approaches have been also applied to VidVRD without substantial modification. However, comparing with VRD in the static image, VidVRD is not only practical but also challenging than VRD, as mentioned in the introduction section. Several well-designed models have been proposed to solve this problem. Shang et al. [20] proposed the first VidVRD framework to temporarily localize and recognize dynamic relationships. they also contributed the first VidVRD dataset which contains rich labeled relations. Tsai et al. [21] proposed a fully-connected spatial-temporal graph constructed for each video and graph convolutional network formulated feature interaction. They proposed constructing a graph similar to the above in a subsequent study but using conditional random fields to take advantage of the statistical dependencies between objects. Sun et al. [11] proposed a video relation model with multi-modal feature fusion and achieved state-of-the-art performance on VidOR dataset in ACM Multimedia 2019 VRU Challenge.

TABLE I
PROPOSED SPATIAL FEATURE CALCULATION

Index	f_1	f_2	f_3	f_4	f_5	f_6
Feature	$\frac{x_{min}+x_{max}}{2}$	$\frac{y_{min}+y_{max}}{2}$	$x_{max}-x_{min}$	$y_{max}-y_{min}$	$\frac{(x_{min}+x_{max})*img_w}{2}$	$\frac{(y_{min}+y_{max})*img_h}{2}$
Index	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
Feature	$(x_{max}-x_{min})*img_w$	$(y_{max}-y_{min})*img_h$	$\frac{x'_{min}+x'_{max}}{2}$	$\frac{y'_{min}+y'_{max}}{2}$	$x'_{max}-x'_{min}$	$y'_{max}-y'_{min}$
Index	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}
Feature	$\frac{(x'_{min}+x'_{max})*img_w}{2}$	$\frac{(y'_{min}+y'_{max})*img_h}{2}$	$(x'_{max}-x'_{min})*img_w$	$(y'_{max}-y'_{min})*img_h$	$\log\frac{h}{h'}$	$\log\frac{h*w}{h'*w'}$

III. THE PROPOSED APPROACH

In the following section, we describe our strategy which is to modify spatial feature extraction and insert skip-connection into FC-layers to improve accuracy for the relation prediction. At first, we describe the proposed spatial feature extraction method in section 3.1. Then, we present the insertion of the proposed skip-connection embedded FC-Layers in section 3.2.

A. Relation Instance Generation

The proposed method is based on a framework which is described in Sun et al. [11] and Shang et al. [20]. It consists of three steps: decomposing one video into segments, predicate recognition on segments and merging relationship predictions in neighboring segments through a greedy association algorithm. We also used pre-computed bounding box trajectories that provided by VRU challenge organizers. We proposed the spatial-temporal feature extraction method that extracts relative location feature and motion feature. We defined the object relative location feature as $f_{Rl} = [f_1, f_2, ..., f_{18}]$. It is calculated as shown in table 1, where (x,y,w,h) and (x',y',w',h') are the bounding box coordinates of subject and object, respectively. (img_w, img_h) is the height and width of the input image. Motion features are defined as follows:

$$f_{Mot} = f^e_{Rl} - f^s_{Rl} \tag{1}$$

This feature extracts various locations over time between the subject and the object, where f^e_{Rl} and f^s_{Rl} are our proposed spatial features extracted from the end and start frames of the candidated segment, respectively. Finally, spatial-temporal features (f_{ST}) are generated by concatenating the features computed above f^e_{Rl}, f^s_{Rl}, and f_{Mot}. We use a pre-trained word2vec model [22], [23] to extract feature f_{Lan} for encoding subject/object categories. It was trained on GoogleNews dataset.

B. Relationship Classification Model

After we generated f_{ST} and f_{Lan}, the features are fed into our two independent classification models which are trained separately. Our model is designed to increase the complexity of the Multi-Layer-Perceptron(MLP) because it is difficult to accommodate the complexity of the input features with a simple MLP model. To this end, we adapted the number of nodes in the MLP and introduced a skip-connection method.

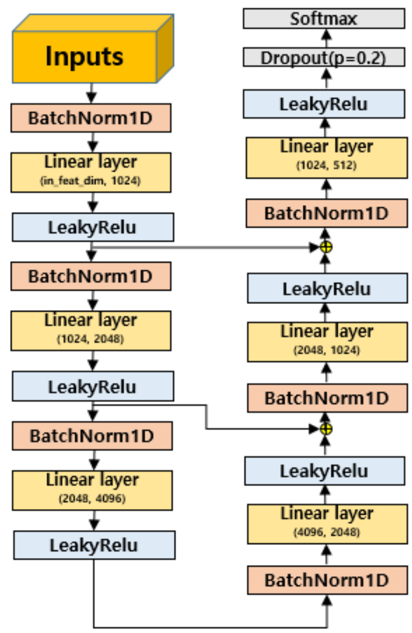

Fig. 3. Proposed relation prediction classifier.

IV. EXPERIMENTS

A. Dataset and Training Details

The VidOR dataset consists of 7,000 videos for train, 835 videos for validation and 2,165 videos for test. 80 categories of objects are annotated with bounding-box trajectory to indicate their spatio-temporal location in the videos; and 50 categories of relation predicates are annotated among all pairs of annotated objects with starting and ending frame index. Our proposed model was trained with Stochastic Gradient Descent (SGD) optimizer, where batch size is 32, momentum is 0.9, weight decay is 0.1, and on NVIDIA GeForce GTX TITAN XP GPU with 12GB memory. The learning rate is set to 0.01, and reduces from 0.01 to 0.0001 for each 10 epochs. The experiments were done with cuDNN v7.5 and CUDA 10.1. for the test , we linearly combine the two prediction confidences of classifiers as follows:

$$P(c_p \mid f_{ST}, f_{Lan}) = \lambda P(c_p \mid f_{ST}) + (1-\lambda) P(c_p \mid f_{Lan}) \tag{2}$$

where c_p denotes prdicate category, λ is set to 0.3.

B. Evaluation Metrics

VRU Challenge adopts Average Precision (AP) to evaluate the detection performance per video and finally calculate

TABLE II
COMPARISON BETWEEN OUR PROPOSED METHOD AND THE TOP-PERFORMING OF THE VRU'19 CHALLENGE ON VIDOR VALIDATION-SET

Method	Tagging precision@1	Tagging precision@5	Tagging precision@10	Recall@50	Recall@100	mAP
Re-produced top-1 solution in VRU'19 challenge	50.48	39.91	32.44	6.69	8.71	6.02
Pre-computed feature + ours model	52.16	**40.35**	**33.03**	6.97	9.08	6.50
Ours feature + model w.o skip-connection	**52.52**	40.18	32.95	6.99	9.12	6.56
Ours	51.68	40.04	33.01	**7.01**	**9.14**	**6.60**

TABLE III
FINAL RESULTS ON VIDOR TESTSET

Method	Tagging precision@1	Tagging precision@5	Recall@50	Recall@100	mAP
Ours	52.69	42.19	7.16	9.36	6.65

the mean AP (mAP) over all testing videos as the ranking score. To match a predicted relation instance $(< s, p, o >^p, (\tau_s^p, \tau_o^p))$ to a ground truth $(< s, p, o >^g, (\tau_s^g, \tau_o^g))$, the requirements should be satisfied as follows: (1) their relation triplets are exactly same, i.e. $< s, p, o >^p = < s, p, o >^g$. (2) $\mathbf{vIoU}(\tau_s^p, \tau_s^g) \geq 0.5$ and $\mathbf{vIoU}(\tau_o^p, \tau_o^g) \geq 0.5$, where $\mathbf{vIoU}$ refers to the volume intersection over union [24]. (3) the minimum overlap of the subject trajectory pair and the object trajectory pair $\mathbf{ov_{pg}} = min(\mathbf{vIoU}(\tau_s^p, \tau_s^g), \mathbf{vIoU}(\tau_o^p, \tau_o^g)$ is the maximum among those paired with the other unmatched ground truths $\mathcal{G}$, i.e. $\mathbf{ov_{pg}} \geq \mathbf{ov_{pg'}}(\mathbf{g'} \in \mathcal{G})$.

C. Results Analysis

Table 3 shows the final results on VidOR test-set. Our proposed method achieves mAP of 6.65%, which is 0.34% higher than the method of the winner in the VRU'19 challenge. We also compared the performances using the VidOR validation set as shown in Table 2 before submitting our final results. The performance is improved when our proposed relationship classification model is connected to the pre-computed features of the VRU'19 challenge winner. Also, the performance of the skip-connection method is slightly enhanced compared to the case of no skip-connection. When both the proposed spatial feature and relationship classification model were applied, there was a better performance improvement than the last year's winning model in most evaluation metrics.

V. CONCLUSION

In this paper, we have proposed a spatial feature extraction and relationship classifier for video visual relation detection in the VidOR dataset. Specifically, the proposed spatial feature extraction method is designed to include the relative position of objects in the image and the relative position between objects. In addition, the relationship classifier is designed to accommodate the complexity of the input features. The experiment results indicate that the proposed model outperforms the last year's winning model in the visual relation detection task of VRU challenge. Our team ($ETRI\_DGRC$) ranked in the 2nd place of the visual relation detection task in the VRU'20 Challenge.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [22ZD1120, Development of ICT Convergence Technology for Daegu-GyeongBuk Regional Industry]

REFERENCES

[1] X. Wu, D. Sahoo, and S. C. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, 2020.
[2] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020.
[3] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128 837–128 868, 2019.
[4] W. Wang, H. Song, S. Zhao, J. Shen, S. Zhao, S. C. Hoi, and H. Ling, "Learning unsupervised video object segmentation through visual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 3064–3074.
[5] Z.-Q. Zhao, S.-T. Xu, D. Liu, W.-D. Tian, and Z.-D. Jiang, "A review of image set classification," *Neurocomputing*, vol. 335, pp. 251–260, 2019.
[6] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
[7] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko, "Joint event detection and description in continuous video streams," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 396–405.
[8] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.
[9] N. Passalis and A. Tefas, "Learning neural bag-of-features for large-scale image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 10, pp. 2641–2652, 2017.
[10] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, "Annotating objects and relations in user-generated videos," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 279–287.
[11] X. Sun, T. Ren, Y. Zi, and G. Wu, "Video visual relation detection via multi-modal feature fusion," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2657–2661.
[12] 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.
[13] 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.

[14] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1385–1392.

[15] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[16] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.

[17] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3038–3046.

[18] 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*, 2016, pp. 379–387.

[19] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, and J. Wang, "Object detection in videos by high quality object linking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 5, pp. 1272–1278, 2019.

[20] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, "Video visual relation detection," in *ACM International Conference on Multimedia*, Mountain View, CA USA, October 2017.

[21] Y.-H. H. Tsai, S. Divvala, L.-P. Morency, R. Salakhutdinov, and A. Farhadi, "Video relationship reasoning using gated spatio-temporal energy graph," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 424–10 433.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[23] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European conference on computer vision*. Springer, 2016, pp. 852–869.

[24] X. Shang, T. Ren, H. Zhang, G. Wu, and T.-S. Chua, "Object trajectory proposal," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 331–336.