Text Description Generation from Videos via Deep Semantic Models

Lin Li* and Kaixi Hu
Wuhan University of Technology, Wuhan, China
E-mail: cathylilin@whut.edu.cn
E-mail: issac_hkx@whut.edu.cn

***Abstract*— Text description generation from videos refers to extracting significant information from a given video, and summarizing them with natural language, which is an effective approach to comprehensively understanding the semantic information hidden in videos. However, conventional videos contain a large number of events where some key information might be reflected in multiple successive events. To this end, this paper focuses on text description generation from videos via deep semantic models that can identify multiple events in a video and capture their dependencies. In particular, we design and propose a series of solutions in terms of scene adaptive video event representation, local event description and video-text summarization generation.**

I. INTRODUCTION

The information of urban people flow, logistics, environmental change and social interaction is recorded in real time by edge devices (e.g., camera) and transmitted by the Internet of Things and cloud services. Surveillance video, as an important form, is characterized by a huge amount of data, diverse content and low value density.

As shown in Figure 1, due to the spatial-temporal correlation and complex semantic information, it is significant to make use of diverse static elements (e.g., objects, figures and backgrounds) in physical space, which is beneficial to capture spatial-temporal variation of visual objects and generate accurate, coherent and logical video-text summarization. To this end, we focus on the problem of multi-view representation learning and semantic modeling in video-text summarization and try to achieve more informative embedding from visual correlation to into linguistic logic.

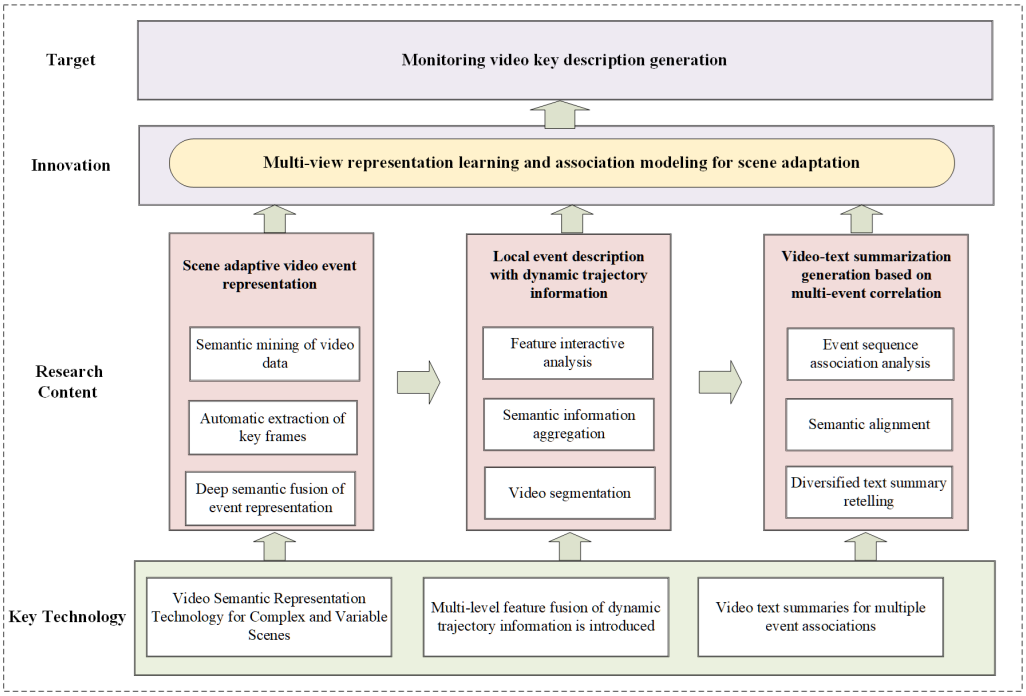

Figure 1. Overall research framework

* Lin Li is the corresponding author.

II. RELATED WORK

Mapping information from different modalities into a joint representation space is a typical multimodal information fusion method, which can be divided into the following three aspects: simple operation based fusion, attention-based fusion and tensor based fusion.

First, the representation of information from different modalities can be fused through simple operations such as splicing and summing, which generally require no or very few parameters and can be adapted to the model structure through joint training with the model. Join operations can be used to fuse low-level input features [1][2][3] or high-level features extracted from pre-trained models [4][5]. Nojavanasghari et al. [1] used viewpoint multimedia (POM) datasets to study persuasiveness and proposed a deep multimodal fusion architecture to predict persuasiveness by linking complementary information from a single modality and quantifying the impact of speakers on listeners' beliefs, attitudes, intentions, motivations, and behaviors. Wang et al. [2] proposed a selective additive learning (SAL) process to improve the accuracy of multimodal emotions by linking different modal information by identifying confounding features in limited data resources. Vielzeuf et al. [4] introduced a central network to connect the networks of specific modalities by assuming that each modality can be handled by a separate deep convolutional network and allowing decisions to be made independently. Zhou et al. [5] used the pre-trained word bag model to connect the word features in the question with the CNN features in the image to predict the answer. For the addition fusion method with weights, an iterative method is proposed in Reference [6], which arranges the pre-trained vectors of the same elements in the order suitable for addition between elements. In addition, the neural structure of progressive exploration [8][9][10] was adopted in the literature [7] to search for appropriate parameters for some fusion functions according to the layer to be fused and the joining or addition operation to be used.

Second, the weight vector dynamically generated by the attention mechanism at each time step is widely used for the fusion of multimodal information [11][12]. Aiming at the attention mechanism of images, literature [13] extended the LSTM model for text processing and added an image attention model conditional on the hidden state of LSTM in the past time step. The input to this model is the connection between the currently embedded word and the participating image features, and the final LSTM hidden state is used as a fused multimodal representation to predict the answer that points to the real VQA. Xu et al. [14] introduced an attention mechanism into the RNN-based encoder-decoder model to allocate attention weights for image features for image description. In addition, literature [15][16] uses the image and query features as conditions to lock the image region related to the answer through the attention mechanism, so as to infer the answer. Different from the above image attention mechanism, collaborative attention mechanism uses symmetric attention structure to generate image feature vectors and language feature vectors [17]. In Literature [18], dual-attention network (DAN) was used to compute the attention distribution of images and texts in parallel under the conditions of features and iteratively updated memory vectors. Stacked hidden attention (SLA) improves SANS [19] by linking the original image representation with the representation of the previous hidden layer in order to preserve the underlying information of the intermediate reasoning stage of attention. Literature [20] establishes the deep correlation between modalities and obtains the attention feature vectors by calculating the inner product between different modal feature representations. As the Transformer model based on self-attention mechanism [21] has achieved good performance in the text field, researchers have extended it to the research of multimodal fusion. The bimodal extension of Bert represents different tokens as a word or an image fragment, and the representations of the image and the word are fused in the input sequence [22][23][24][25][26].

Thirdly, bilinear convergence is a common method for the fusion of image feature vectors and text feature vectors. This method creates joint representation space by calculating the cross product of different modal representations to facilitate the multiplication interaction between all elements in the two vectors. This method is also called second-order pooling [27]. The bilinear representation usually uses the linear transformation of the two-weight value matrix to the output vector. When calculating the cross product, each eigenvector can be extended by an additional value, so that the single-modality input feature is maintained in the bilinear representation [28]. On the one hand, due to the correlation between bilinear representation and the kernel in polynomials, a more dense representation [29] can be obtained through different degrees of low-dimensional approximations, such as calculation of Sketch[30], convolution [31], low-rank decomposition [32], etc. On the other hand, bilinear convergence can also be combined with the Attention mechanism. Kim et al. [32] used the textual representations of multimodal low-rank bilinear pooling (MLB) as the input of the Attention model to obtain the attention-processed image representations, and then used the MLB model to integrate the textual representations. You get a joint representation.

III. SCENE ADAPTIVE VIDEO EVENT REPRESENTATION

As shown in Figure 2, the semantic mining technology of surveillance video data is studied by employing deep clustering analysis of semantic related video frames. This technology can deal with the problem of repeat information in continuous frames in input videos and make full use of surveillance video information. According to the classification difference between video frames in surveillance video, the automatic extraction method of scene adaptive key frames in surveillance video is studied based on scene knowledge transfer, which reduces the interference of redundant information and improves the discriminant ability

of effective information. Aiming at the objects, behaviors and scenes in video frames, the high-level semantic information in the rich visual features of surveillance video sequences is mined, and the scene adaptive video event representation method is studied by means of collaborative learning.

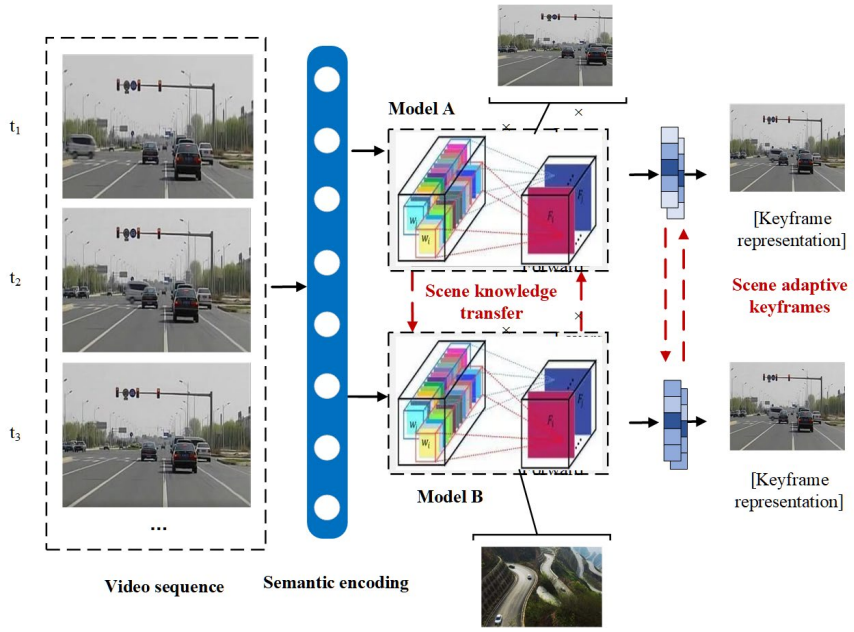

Figure 2. Scenario adaptive event representation

IV. LOCAL EVENT DESCRIPTION

As shown in Figure 3, focusing on the rich image sequence information in multi-source surveillance video, combined with target detection and depth feature extraction technology, interactive analysis of context feature information is realized. Through the multi-view representation learning technology, the deep semantic information aggregation method based on multi-level features is studied according to semantic segmentation technology and feature weight learning on event dimensions. In this paper, a video segment segmentation method based on sparse annotation is studied. The location and category information are used to mine the regional feature information of video clips. Aiming at the dynamic moving trajectory data of the target object, the fusion and complementation of static frame information and dynamic trajectory information is realized. The method of multi-target tracking is used to study the local event description method in surveillance video based on multi-level feature interactive analysis in the way of self-supervised learning.

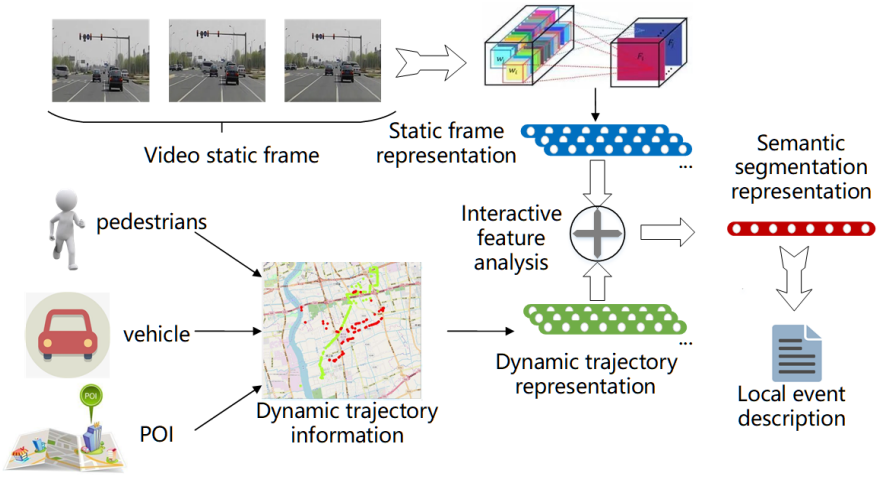

Figure 3. Local event description with dynamic trajectory information

V. VIDEO-TEXT SUMMARIZATION GENERATION

As shown in Figure 4, in view of the spatial-temporal information of surveillance video data, the correlation analysis method of context event sequence is studied. Moreover, the method of semantic alignment between visual features and text is focused. And cross-modal retrieval is used to dynamically search the retrieval sentences related to video in corpus, which provides guidance for subsequent generation of video content description. Considering accurate, diverse, controllable and coherent, paraphrase of video summarization is studied by using multi-objective deep reinforcement learning method. In particular, syntactic controlled paraphrase based on multi-task learning can be achieved by introducing syntactic information. And, Automatic text description generation based on event multi-sentence description is designed by combining events development rules and text language characters.

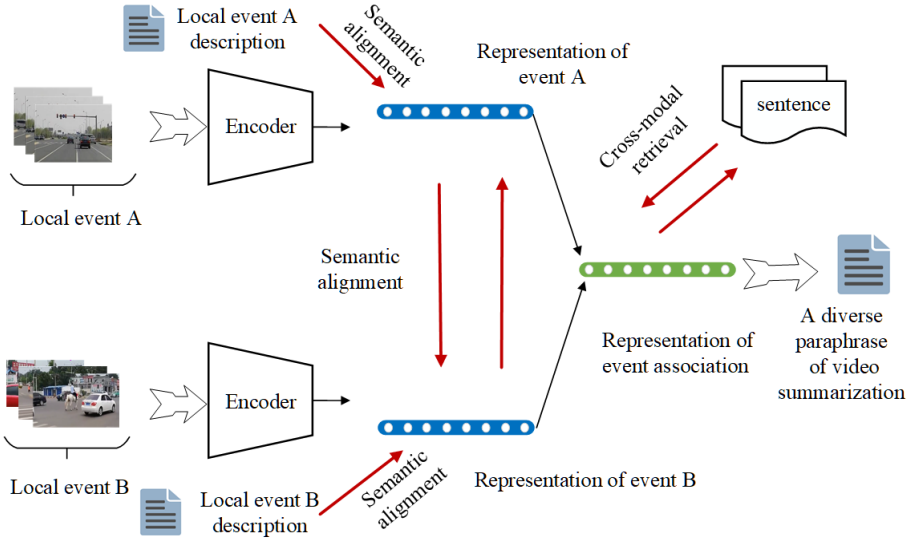

Figure 4. Video-text summarization generation associated with multiple events

VI. CONCLUSIONS

Based on the videos generated by monitoring equipment, this paper discusses a solution about text description generation from videos by multi-event association and semantic analysis which covers scene adaptive video event representation, local event description and video-text summarization generation. This solution can assist government management, decision-making and improve the service of smart city.

REFERENCES

[1] B. Nojavanasghari, D. Gopinath, J. Koushik, B. T., and L.-P. Morency,"Deep multimodal fusion for persuasiveness prediction," in Proc. ICMI, 2016.
[2] H. Wang, A. Meghawat, L.-P. Morency, and E. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in Proc. ICME, 2017.
[3] A. Anastasopoulos, S. Kumar, and H. Liao, "Neural language modeling with visual features," in arXiv: 1903.02930, 2019.
[4] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in Proc. ECCV, 2018.
[5] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," in arXiv:1512.02167, 2015.
[6] J.-M. P´ erez-R´ ua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie,"MFAS: Multimodal fusion architecture search," in Proc. CVPR, 2019.
[7] B. Zoph and Q. Le, "Neural architecture search with reinforcement learning," in Proc. ICLR, 2017.
[8] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, F.-F. Li, A. Y uille, J. Huang, and K. Murphy, "Progressive neural architecture search," in Proc. ECCV, 2018.
[9] J.-M. P´ erez-R´ ua, M. Baccouche, and S. Pateux, "Efficient progressive neural architecture search," in Proc. BMVC, 2019.
[10] X. Y ang, P . Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in Proc. ACM MM, 2016.
[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.
[12] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," in arXiv: 1410.5401, 2014.
[13] Y. Zhu, O. Groth, M. Bernstein, and F.-F. Li, "Visual7W: Grounded question answering in images," in Proc. CVPR, 2016.
[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y . Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proc. ICML, 2015.
[15] K. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in Proc. CVPR, 2016. 152
[16] Z. Y ang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in Proc. CVPR, 2016.

[17] J. Lu, J. Y ang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in Proc. NIPS, 2016.
[18] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in Proc. CVPR, 2017.
[19] H. Fan and J. Zhou, “Stacked latent attention for multimodal reasoning,” in Proc. CVPR, 2018.
[20] I. Schwartz, A. Schwing, and T. Hazan, “High-order attention models for visual question answering,” in Proc. NIPS, 2017.
[21] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.
[22] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, “Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training,” in arXiv: 1908.06066, 2019.
[23] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, “VL-BERT: Pre-training of generic visuallinguistic representations,” in arXiv: 1908.08530, 2019.
[24] L. Li, M. Y atskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visual-BERT: A simple and performant baseline for vision and language,” in arXiv: 1908.03557, 2019.
[25] C. Sun, A. Myers, C. V ondrick, K. Murphy, and C. Schmid,“VideoBERT: A joint model for video and language representation learning,” in Proc. ICCV, 2019.
[26] C. Alberti, J. Ling, M. Collins, and D. Reitter, “Fusion of detected objects in text for visual question answering,” in Proc. ICMLC, 2019.
[27] J. Tenenbaum and W. Freeman, “Separating style and content with bilinear models,” Neural Computing, vol. 12, pp. 1247–1283, 2000.
[28] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in Proc. EMNLP, 2017.
[29] Y . Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in Proc. CVPR, 2016.
[30] M. Charikar, K. Chen, and M. Farach-Colton, “Finding frequent items in data streams,” in Proc. ICALP, 2012.
[31] N. Pham and R. Pagh, “Fast and scalable polynomial kernels via explicit feature maps,” in Proc. SIGKDD, 2013.
[32] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang,“Hadamard product for low-rank bilinear pooling,” in Proc. ICLR,2017.