MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
http://www.comsoc.org/~mmc

MMTC Communications - Frontiers

Vol. 15, No. 6, November 2020

CONTENTS

SPECIAL ISSUE ON the Development and Innovation of Information and Communication Technology

Guest Editor: Bin Tan
Jinggangshan University, China
Email: tanbin@jgsu.edu.cn

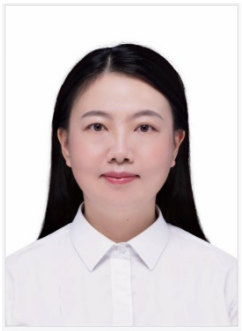
This special issue of Frontiers focuses on the development and innovation of Information and Communication Technology field. The research topics of the papers in this special issue include how to introduce attention mechanism, bilinear pooling, and depth wise separable convolution to the defect detection model, a new 360° VR system over C-RAN, translating video into language by enhancing visual and language representations, and optimize the energy efficiency of video analytics tasks using a variable-resolution strategy.

The first paper proposed a novel training strategy to form a new object-level attention mechanism, and construct an efficient CNN model based on the object-level attention mechanism and bilinear pooling for the DR defect inspection. The defects of the casting on radiography images were effectively recognized in complicated detection scenario. The proposed model outperformed other classical deep learning classification models in each quantitative metric.

The second paper design a view synthesis-based 360° VR caching system over C-RAN, where MEC is enabled for view synthesizing and hierarchical caches are deployed at both BBU pool and RRHs. To decrease the transmission latency and backhaul traffic load for VR services, an integer linear program problem is formulated.

The third paper presented a two-stage pre-training strategy which is more effective for video description. The used CNN model is first pre-trained on Imagenet to prevent the model from falling into over-fitting. And then, the model is fine-tuned on MSCOCO, having the model sensitive to frequently used words and sentence patterns in the task of video captioning. A visual and language representation enhancing method is proposed to capture more comprehensive information including static and motion visual feature and personalized language feature. Additionally, the proposed enhancing method is applied on both un-factored way and factored way LSTM networks for video description.

The fourth paper investigated the ability to perform intelligent video analytics in energy-constrained edge devices. To globally optimize energy efficiency, their RL network learns the best non-myopic policy for determining the spatio-temporal frame resolution of incoming video stream data. Compared with other energy-efficient single-task video analytics solutions that were designed for still images without utilizing temporal information, their work is the first to address the energy consumption optimization problem for multi-task video analytic pipeline, and it is also the first to leverage RL to holistically tackle all these challenges indicated above, and to do end-to-end global efficiency policy optimization.



Bin Tan received the B.S. degree from the Jiangxi University of Science and Technology, Ganzhou, China, in 2003, the M.S. degree from Wuhan University, Wuhan, China, in 2008, and the Ph.D. degree in computer science and technology from Tongji University, Shanghai, China, in 2017.She is currently an Associate Professor with the College of Electronics and Information Engineering, Jinggangshan University, Ji'an, China. Her current research interests include multimedia communication, source and channel coding, and wireless networking.

An Efficient Convolutional Neural Network Model Based on Object-Level Attention Mechanism for Casting Defect Detection on Radiography Images

Chuanfei Hu and Yongxiong Wang
School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology
w64228013@126.com; wyxiong@usst.edu.cn

1. Introduction

Aluminum castings have been extensively applied to various parts whose qualities affect the fatigue behavior of the overall product [1]. Due to the complexity and diversity of the casting process [2], defects are inevitable in the internal of castings, such as gas holes, sand holes, and flaws. In order to obtain the internal in formation and guarantee the completeness of castings, radiography is often used for nondestructive testing [3], which has been widely used in quality controlling [4]. In radiographic testing processing, most manufacturers rely mainly on manual detection according to the experiences of operators about the shape, brightness, and contrast of castings. Such a manual method is not only mechanical and inefficient, but also may cause ophthalmic diseases due to the high-frequency illumination of the display. Consequently, the digital radiography (DR)-based automatic inspection system has been one of the research focuses.

Recently, deep learning model, particularly the convolutional neural network (CNN), has received substantial interest in industrial applications [5]–[9]. In DR defect inspection, which requires quick and reliable results, the powerful deep learning model may not be the best solution due to its huge computation and high dependence of precise annotations. It is a main issue to make a good tradeoff between the performance and efficiency of execution in a real-time system. For another problem, labeling each defect in each radiography image is very expensive and laborsome. An alternative way is to annotate the images with coarse labels only indicating whether there are any defects in the image or not. However, an image-level label does not have enough position information of the defects. It is not easy for the classification model to extract defect features directly, because the differences between defect and non-defect images include not only defects, but also the complex structures of castings and background in the complicated scenario.

To overcome the challenges of subtle defect representations and lack of precise annotations, we introduce attention mechanism, bilinear pooling, and depth wise separable convolution to the defect detection model. The main contributions are as follows: 1) A new CNN model, including type classification module (TCM) and defect classification module (DCM), is constructed for DR defect inspection. Simultaneously, bilinear pooling is used to obtain strong feature representation, and depthwise separable convolution is introduced in DCM to reduce computation. In real-time complex application, the proposed model outperforms classical deep classification models in terms of each quantitative metric (e.g., accuracy, precision, recall, F-measure, FPS, and Para Size). 2) Compared with previous methods of attention mechanism, a novel training strategy is proposed to construct object-level attention mechanism which is effective and free of computing burdens. To the best our knowledge, we are among the first to achieve the attention mechanism without additional architectures for defect detection models. 3) Inspired by class activation maps (CAM) [9], we propose bilinear CAM (Bi-CAM), which is suitable to bilinear architectures, to be used as a visualization technique to increase the interpretability of the model.

2. Strategy for Object-Level Attention Mechanism

In our task, any spatial information of defects cannot be provided because the dataset is only annotated with image-level labels, defect or non-defect. Thus, attention mechanism is an indispensable tool in complicated scenarios. In order to guarantee an industrial inspection in real time, we propose a novel training strategy without additional network structures to form an attention mechanism, namely, object-level attention mechanism. The basic idea of the scheme is motivated by the process of teaching infants, where parents often teach infants to pay more attention to recognizing the low-level objects first, such as fruits and bottles, then infants are told whether they can eat these objects. Inspired by this way, the proposed training strategy is to let the model pay attention to the certain type of the casting in the image, and then, the model can be taught to infer defects accurately based on the cognition. In the first stage of the proposed strategy, the object-level attention mechanism which bridges the gap between image-level annotations and application scenario is generated.

According to the proposed strategy, we set two datasets with type and defect labels, respectively. Simultaneously, a novel CNN model is constructed where the two subnetworks are cascaded, namely, TCM and DCM. During the strategy, we first train TCM on the type dataset with an additional classifier to

distinguish the type of castings, then the softmax layer is discarded after TCM converges. In the secondary stage, the total model is trained on the defect dataset while the parameters of TCM are fixed. The object-level attention mechanism enables TCM to provide object-related features for subsequent DCM. Overall framework

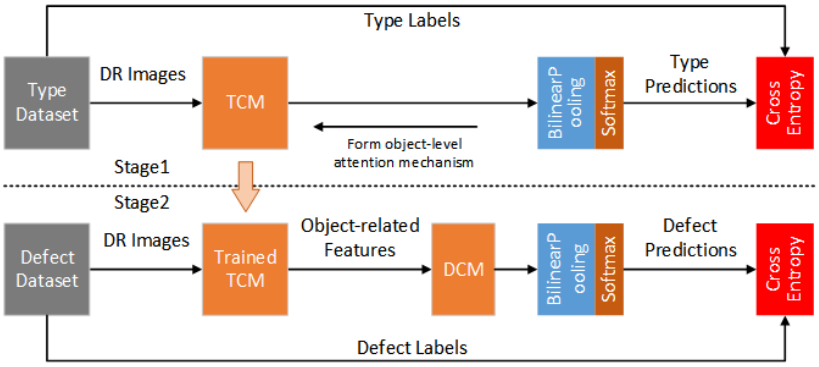


Figure 1 Overall framework of the proposed method.

of the method is depicted in Figure 1.

3. Type Classification Module

TCM is in charge of irrelevant suppression and object-related representation in complex scenario via the object-level attention mechanism. We implement the TCM based on VGG16 [9] which is effective and light. Given an input image, the feature maps are extracted by TCM which can be formulated as:

$$f_{TCM}(X; w_{TCM}, b_{TCM}) = w_{TCM} * X + b_{TCM}$$

where X is the input radiography image, w_{TCM} and b_{TCM} are the parameters of convolutional layers in TCM which can be learned, $*$ represents the convolutional operation. $f_{TCM}(X; w_{TCM}, b_{TCM})$ is the output feature maps of TCM.

To improve the representation of subtle defects, the bilinear pooling is inserted before prediction layer. Assuming that the output of the module is $X \in R^{L \times D}$, $Y = X$, where L and D are the number and channel of features, respectively. Then, the bilinear aggregation is denoted as:

$$Z = X^T Y$$

where Z is the result of matrix multiplication for X and Y.

The output nodes of the last layer are the same as the number of type classes. The softmax activation function can be regarded as an estimation of probability of each class since the sum of softmax classifier is one. We consider the class with the highest probability as the final result from the classifier. In the training stage, the cross-entropy loss function is leverage to measure the error between the prediction and the target label.

4. Defect Classification Module

DCM aims at achieving a tiny defect classification by mining deeper features in the object-related features from TCM. To meet industrial inspection requirements of accuracy, speed, and size of model, we attempt to design DCM by stacking few computational units which are efficient and similar with residual blocks. Due to the efficiency of depthwise separable convolution [10], we replace entire standard convolutions to the depthwise separable convolutions in residual blocks. The data stream of DCM can be formulated as:

$$f_{DCM}(X; w_{DCM}, b_{DCM}) = w_{DCM} \otimes f_{TCM}(X) + b_{DCM}$$

where $f_{TCM}(X)$ are the feature maps from TCM, $\mathrm{w_{DCM}}$ and $\mathrm{b_{DCM}}$ are learnable parameters of depthwise separable convolutional layers in DCM. $\otimes$ represents the depthwise separable convolutional operation. Before the softmax layer, the bilinear pooling is also used to enhance the representation of the subtle defects. We still use cross-entropy loss to train DCM in secondary training stage.

5. Bilinear Class Activation Maps

In order to interpret the validity of the proposed model visibly, we require a visual method to depict the relationship between feature maps and predictions. Inspired by CAM [8], we propose a new method, called Bi-CAM, to generate the active feature map for bilinear architectures by the simple algorithm. The procedure for generating CAM $M_c(x, y)$ of class c is formulated by:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y)$$

where $f_k(x, y)$ represents the feature map of channel k in the last convolutional layer of the classification

network at spatial (x, y). w_k^c is the weight corresponding to class c in the softmax layer for channel k. However, the original procedure is only suitable to global average pooling (GAP) architectures. To adapt bilinear architectures in our model, we extend the CAM method. The bilinear pooling is instanced such that

$$Z = X^T Y = X^T X = \begin{pmatrix} f_1^T f_1 & \cdots & f_1^T f_D \\ \vdots & \ddots & \vdots \\ f_D^T f_1 & \cdots & f_D^T f_D \end{pmatrix}$$

where Z is a Gram matrix associated with feature map f_k, $k = 1, \ldots, D$, where f_k is an abbreviation of $f_k(x, y)$, $z_{ij} = f_i^T f_j$ is the outer product of f_i and f_j, $i, j \in k$. Following [9], but being quite different, the score of class c in the softmax layer is denoted by:

$$S_c = \sum_{i,j \in k} w_{ij}^c f_i^T f_j = \sum_{i,j \in k} w_{ij}^c z_{ij}$$

where w_{ij}^c is the weight corresponding to z_{ij}. To spotlight the active regions of predicted class, a set of weights is required to describe the activation of each feature map. Thus, we first construct a square matrix consisting of w_{ij}^c. Second, eigenvalues $\widetilde{w}_{ii}^c$ are obtained by eigen-decomposition approach. $\widetilde{w}_{ii}^c$ can approximate the importance of f_i, since more information of each channel is concentrated in the corresponding eigenvalue. Finally, Bi-CAM is formulated by:

$$M_c^{bi} = \sum_{i \in k} \widetilde{w}_{ii}^c f_i.$$

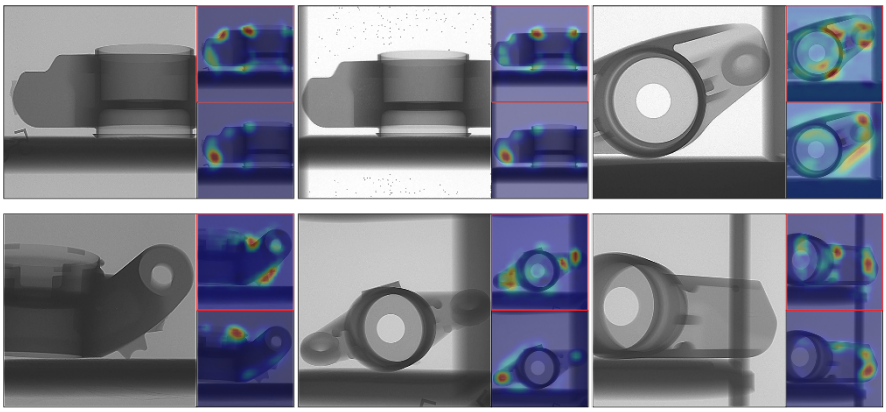


Figure 2 Some examples of CAM and Bi-CAM, where the images based on Bi-CAM are labeled with the red rectangles.

6. Experimental results

The radiography images, provided by the cooperative enterprise, are generated by the inspection equipment named Y.MU2000-D which is an industrial X-ray and CT inspection system produced by YXLON. In order to adapt our proposed training strategy, the dataset is divided into type and defect datasets which are independent, respectively. The type dataset includes 9215 images of 51 categories for each type of castings. TCM is trained on the training set which randomly selected 70% images from each category of the type dataset. We evaluate the module on the testing set on the remaining images. As defect images produced by the system are rare, the defect dataset includes 1469 images of 20 categories only with image-level defect labels. We alter the distribution of training set and testing set to 60% and 40% in the defect dataset in order to verify the generalization ability of the model.

In Table 1, our model is superior than VGG16 and ResNet50 in terms of each quantitative evaluation metric. Then, we further visualize qualitative results to describe the differences between CAM and Bi-CAM. As shown in Figure 2, Bi-CAM surely highlights more details of castings than CAM.

Table 1 Comparisons with VGG16 and ResNet50.

Model	Accuracy (%)	Precision (%)	Recall (%)	F-M (%)	FPS	Para Size (MB)
VGG16	54.51	45.51	57.08	50.65	40.67	1688.19
ResNet50	56.90	46.52	36.25	40.75	41.07	89.99
Ours	**92.77**	**90.67**	**91.60**	**91.31**	**49.11**	**58.27**

7. Conclusion and future

In this letter, we proposed a novel training strategy to form a new object-level attention mechanism, and construct an efficient CNN model based on the object-level attention mechanism and bilinear pooling for the DR defect inspection. The defects of the casting on radiography images were effectively recognized in complicated detection scenario. The proposed model outperformed other classical deep learning

classification models in each quantitative metric. It demonstrated that our model has a reliable advantage in efficiency and can be used for a real-time DR inspection system with complicated background.

Researchers should make the best use of training strategies or network architectures to improve the performance and reduce additional labor consumption, such as manual annotation. This is a trend of designing industrial detection model. However, compared to direct supervision methods, the drawback is that our model cannot predict excessive results, such as the positions and types of defects. In future work, we aim to investigate topology analysis to infer a coarse location by the correlation of features and utilize metric learning to divide the types of defects by the learnable model.

References

[1] H. Mayer, "Recent developments in ultrasonic fatigue," Fatigue Fracture Eng. Mater. Struct., vol. 39, no. 1, pp. 3–29, Jan. 2016.

[2] S. Pattnaik, D. B. Karunakar, and P. K. Jha, "Developments in investment casting process—A review," J. Mater. Process. Technol., vol. 212, no. 11, pp. 2332–2348, Nov. 2012.

[3] Y. Hangai et al., "Nondestructive observation of pore structure deformation behavior of functionally graded aluminum foam by X-ray computed tomography," Mater. Sci. Eng., A, vol. 556, pp. 678–684, Oct. 2012.

[4] J. Wang, P. Fu, and R. X. Gao, "Machine vision intelligence for product defect inspection based on deep learning and hough transform," J. Manuf. Syst., vol. 51, pp. 52–60, 2019.

[5] M. S. Hossain,M. H. Al-Hammadi, and G. Muhammad, "Automatic fruits classification using deep learning for industrial applications," IEEE Trans. Ind. Informat., vol. 15, no. 2, pp. 1027–1034, Feb. 2019.

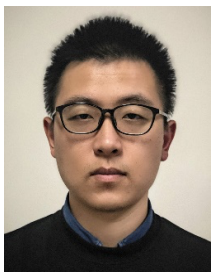
[6] Q. Xuan, Z. Chen, Y. Liu, H. Huang, G. Bao, and D. Zhang, "Multiview generative adversarial network and its application in pearl classification," IEEE Trans. Ind. Electron., vol. 66, no. 10, pp. 8244–8252, Oct. 2019.

[7] Z. Zhong and Z. Ma, "A Novel Defect Detection Algorithm for Flexible Integrated Circuit Package Substrates," IEEE Trans. Ind. Electron., vol. 69, no. 2, pp. 2117-2126, Feb. 2022.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 2921–2929.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1800–1807.



Chuanfei Hu received the B.S. degree in electrical engineering and automation from the Jiangsu University of Science and Technology, Zhenjiang, China. and the M.S. degree in control engineering from the University of Shanghai for Science and Technology, Shanghai, China. He is currently working toward the Ph.D. degree in electronic information toward the Southeast University, Nanjing, China. His research interests include computer vision and applications of deep learning.

Yongxiong Wang received the B.S. degree in engineering mechanics from Harbin Engineering University, Harbin, China, and the M.S. and Ph.D. degrees in control science and engineering from Shanghai JiaoTong University, Shanghai, China. He is currently a Professor of Control Science and Engineering with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include computer vision and intelligent robot.

A View Synthesis-based 360° VR Caching System over MEC-enabled C-RAN

Jianmei Dai, Zhilong Zhang, Shiwen Mao and Danpu Liu
Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, China
Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA
{jammy_d-ane, dpliu }@bupt.edu.cn; zhilong.zhang@outlook.com; smao@ieee.org

1. Introduction

Virtual reality (VR) is a human computer interface technology that enables users to interact with each other in the virtual environment with three-dimensional spatial information [1]. A recent market report forecasts that the data consumption from mobile VR devices (smartphone-based or standalone) will grow by over 650% between 2017 to 2021 [2], [3]. 360 video is an integral part of VR. As a user can freely change his/her viewing direction while watching, it can provide panoramic and immerse experience. Nevertheless, wireless 360 video delivering incurs 4-5 times higher bandwidth requirements than that of traditional videos. Research by Huawei iLab shows that a general 360° VR video data rate with 4K resolution is 50 Mbps, and the data rate with 8K resolution increases to 200 Mbps [4]. Therefore, with the rapid increase in the number of VR headsets (wireless VR headsets are going to increase to 50 million by 2021) [2], the communication network can potentially become a bottleneck.

Some VR solutions use user's field of view (FoV) streaming to reduce bandwidth consumption [5]–[9].While FoV adaptive 360 video streaming is useful for reducing bandwidth requirements, 360 video streaming from remote content servers is still challenging due to network latency. The Latency restriction is critical for VR services. Many studies indicate that the motion-to-photon (MTP) latency for VR should be less than 20 ms; otherwise, the user will feel dizzy. To alleviate the transmission latency, an efficient approach is caching popular VR contents at the edge of the network, such as RRHs and base stations. The existing literature has studied a number of problems related to caching in VR systems [10]–[13]. However, these caching schemes do not take the view synthesis character into consideration. View synthesis is a feature of multi-view video. Many methods [14]–[17] can be used for view synthesis, for example, Depth-Image-Based Rendering (DIBR) [14] technique, which is the most widely used method, can synthetically generate free-viewpoint video by using a reference 2D video and its associated depth map.

View synthesis is not only a common way to generate free-viewpoint video from a limited number of views, but also an effective method for predictive coding in multi-view video compression [18]–[20] and can achieve good performance. Moreover, view synthesis can be utilized in VR systems to generate corresponding views according to the viewpoint of users [21]. Indeed, the user's current desired FoV can be synthesized by the previous requested nearby FoVs in the 360° VR video streaming, because the adjacent FoVs usually share many similar parts. Based on this, if a required FoV which can be used for synthesizing more incoming FoVs is cached, the user requests can be largely satisfied by transmitting only a part of FoVs (correspondingly, a part of tiles) from the source. Therefore, we propose a new 360° VR system over C-RAN, where both mobile edge computing (MEC) and hierarchical caching are supported. Therefore, the transmission latency and backhaul traffic load for 360° VR services can be decreased, and the energy consumption on mobile phones is significantly relaxed. Different from [22], *i) in the proposed VR system, the video data do not need to be pre-fetched, so there are no additional remote access cost and local transmission cost; ii) the caching is hierarchical and cooperative, which is more suitable for the C-RAN architecture and iii) The view synthesis is done by MEC-Cache server in the BBU pool, due to the ample computing resource, the processing latency is less than the smartphones, which can significantly increase the QoE of VR users.*

Furthermore, to fully exploit the benefits of the proposed view synthesis-based 360° VR caching system, several challenges need be addressed. First, view synthesis is a computationally intensive task. The concurrent video synthesis could quickly exhaust the available processing resources of the MEC-Cache Server. Therefore, an efficient cache scheme needs to be designed for the given processing resources. Second, caching multiple views of video incurs high overhead in storage. Although hard disks are now very cheap, storing all these files is neither economical nor feasible. Finally, the impact of caching data at the BBU pool and at different RRHs should be quantified, and the questions of what contents and where to be placed should be addressed.

2. System Model

We show the view synthesis-based 360° VR caching system over C-RAN in Fig.1, which consists of one BBU

pool and a set of RRHs connected to the BBU pool via low-latency, high-bandwidth fronthaul links. An MEC-Cache server is deployed at the BBU pool, providing computing, synthesizing, caching, and networking capabilities to support context-aware and delay-sensitive applications near the users. The MEC-Cache server and RRH are with caching capabilities. A set of 360° VR video files are stored in the VR video source server, which can be transmitted and cached in the C-RAN network.

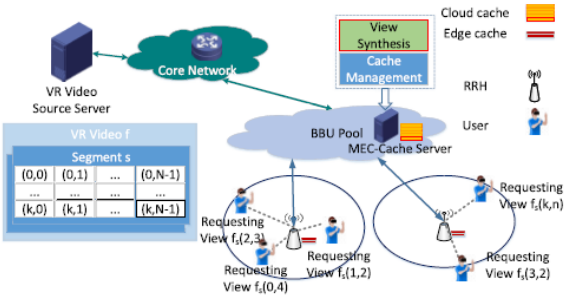


Fig. 1 Illustration of the view synthesis-based 360° VR caching system over Cloud Radio Access Network.

We consider that video requests arriving at each RRH following a Poisson process. The caching design is evaluated in a long time period to accumulate many request arrivals. Basically, one user only connects to one nearest RRH (in terms of signal strength) at the same time, which is later referred to as the user's associated RRH. The data can be fetched either from the associated RRH cache or from the BBU pool, to offload the traffic and reduce the transmission delay of both the fronthaul and backhaul. If the required view is not cached, it can be synthesized by the MEC-Cache Server since the left and right views are cached in the BBU pool. Furthermore, considering that many users mainly download and watch videos with little data uploading to the VR video source server, and the uplink of the fronthaul is idle most of the time, the requested view data can be obtained and synthesized from other unassociated RRH caches. This method can reduce the consumption of backhaul resource much further. Otherwise, the users should obtain the requested view data from the video VR source server.

3. Hierarchical Video Caching Algorithm

We illustrate seven possible (exclusive) events that happen when a user is requesting video view data, and the total download latency in the network for a user request is computed. To minimize the overall downloading latency in the network, we formulate an optimal problem. The problem is NP-Hardness and is extremely challenging to solve in polynomial time. Therefore, a brief analysis of an optimal solution to serve as a performance baseline is illustrated, and an online view synthesis-based caching algorithm is proposed.

The whole process of the caching process is as follows: the requested view data is checked at the beginning of every process loop, if the data can be fetched from the BBU pool or the associated/unassociated RRH caches or can be synthesized by the MEC-Cache server, it will be sent to the user immediately. Otherwise, the system will bring it to the user from the VR video source server. In the meantime, the caching stage is launched. There are two phases in the caching stage. One is the cache placement phase, in which the data is cached immediately since the cache storage is not full; the other is the cache replacement phase, in which the cache storage is full. In the cache replacement phase, if the video segment is new, which means that no views of one segment in a video are cached at the RRHs or the BBU pool, the least frequently used data will be replaced. If not, considering the view synthesis feature of 360° VR data, we use the proposed MaxMinDistance (hereinafter referred to as MMD) scheme to replace the cached data.

The main idea of MMD scheme is to get the smallest maximum distance for all segments. We prove that the proposed online view synthesis-based caching algorithm has a competitive ratio of 2, compared with the optimal offline algorithm for solving the minimization optimization problem. We also prove that it is an efficient/easy algorithm for common three-dimension VR videos and the polynomial time is needed. Furthermore, in terms of space, it will consume only linear space complexity which is nothing but size of given elements.

4. Simulation Results

We perform numerical simulations to evaluate the performance of the proposed MMD algorithm. Five existing algorithms, including the KP-optimal algorithm, the traditional LFU algorithm, VS-RANDOM algorithm, VS-LFU algorithm and Efficient View Exploration Algorithm (EVEA) [22], are compared with the proposed scheme. The performance metrics used for the evaluation are Average cache hit rate (AHR), Backhaul traffic load, Average latency, and Quality-of-experience (QoE).

Firstly, we change the total cache size from 160 GB (10% of the total file size) to 480 GB (30% of the total file size) to observe the performance in terms of the cache hit rate, backhaul traffic load, average latency and QoE. As shown in Fig.2, the performance of the six algorithms improves as the cache size grows, and the MMD algorithms always achieve an obviously superior performance, which is benefit from considering view synthesis feature in caching and processing.

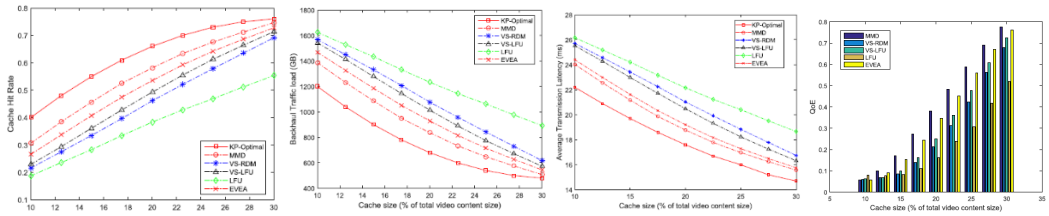


Fig. 2 Impact of Cache Storage.

Secondly, knowing that the quality of each synthesized view depends on its distance to its two reference views, we fix the cache capacity to 160 GB and change the view synthesize range from default 2 to 10 to observe the influence on the performance in this subsection. From Fig.3, we can see that the larger the view synthesis range, the more obvious the advantage of this algorithm is, in addition to the KP-optimal algorithm. Further, while the average video quality will decrease when the view synthesis range increases, the rebuffed time and startup delay will decrease sharply. Therefore, we can see that the QoE remains increasing as the view synthesis range increases and the performance of MMD algorithm is the best among all imported schemes.

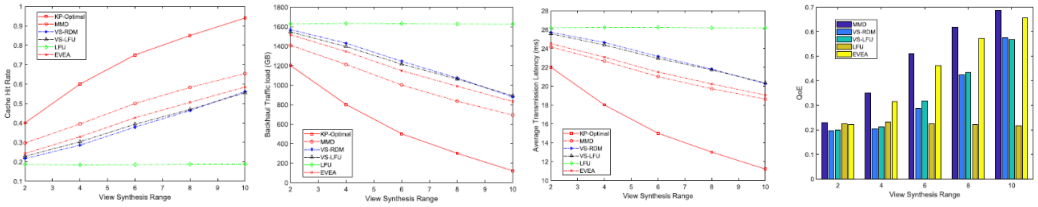


Fig. 3 Impact of View Synthesis Range.

5. Summary

In this e-letter, we design a view synthesis-based 360° VR caching system over C-RAN, where MEC is enabled for view synthesizing and hierarchical caches are deployed at both BBU pool and RRHs. To decrease the transmission latency and backhaul traffic load for VR services, an integer linear program problem is formulated. Due to the NP-completeness of the problem and the absence of the request arrival information in practice, we propose an efficient online MMD caching replacement algorithm, which is proved sub-optimal and low complexity. Rigorous numerical simulations show that the proposed algorithm always yields better performance in terms of cache hit rate, backhaul traffic load, average transmission latency and QoE than the other employed caching algorithms.

References

[1] N. S. S. Hamid, F. A. Aziz, and A. Azizi, “Virtual reality applications in manufacturing system,” in Proc. Sci. Inf. Conf., Aug. 2014, pp. 1034–1037.

[2] Forecast ‘Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021’, Cisco, San Jose, CA, USA, 2017.

[3] Juniper. (2017). Virtual Reality Markets: Hardware, Content & Accessories 2017–2022. [Online]. Available: https://www.juniperresearch.com/researchstore/innovationdisruption/virtual-reality/hardwarecontent-accessories

[4] Huawei. (2016). White Paper: Hosted Network Requirements Oriented on VR Service. [Online]. Available: http://www.imxdata.com/archives/17346

[5] M. Hosseini and V. Swaminathan, “Adaptive 360 VR video streaming:Divide and conquer,” in Proc. IEEE Int. Symp. Multimedia (ISM), Dec. 2016, pp. 107–110.

[6] M. Graf, C. Timmerer, and C. Mueller, “Towards bandwidth efficient adaptive streaming of omnidirectional video over HTTP: Design, implementation, and evaluation,” in Proc. 8th ACM Multimedia Syst. Conf., 2017, pp. 261–271

[7] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, “Hevccompliant tile-based streaming of panoramic video for virtual reality applications,” in Proc. 24th ACM Int. Conf. Multimedia, 2016, pp. 601–605.

[8] R. Skupin, Y. Sanchez, D. Podborski, C. Hellge, and T. Schierl, “HEVC tile based streaming to head mounted displays,” in Proc. 14th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC), Jan. 2017, pp. 613–615.

[9] D. V. Nguyen, H. T. T. Tran, A. T. Pham, and T. C. Thang, “An optimal tile-based approach for viewport-adaptive 360-degree video streaming,” IEEE J. Emerg. Sel. Topics Circuits Syst., vol. 9, no. 1, pp. 29–42,Mar. 2019.

[10] S. Sukhmani, M. Sadeghi, M. Erol-Kantarci, and A. El Saddik, "Edge caching and computing in 5G for mobile AR/VR and tactile Internet," IEEE Multimedia Mag., vol. 26, no. 1, pp. 21–30, Jan./Mar. 2019.

[11] J. Chakareski, "VR/AR immersive communication: Caching, edge computing, and transmission trade-offs," in Proc. Workshop Virtual Reality Augmented Reality Netw., Aug. 2017, pp. 36–41.

[12] M. Chen, W. Saad, and C. Yin, "Echo-liquid state deep learning for 360content transmission and caching in wireless VR networks with cellular-connected UAVs," 2018, arXiv:1804.03284. [Online]. Available: https://arxiv.org/abs/1804.03284

[13] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Edge computing meets millimeter-wave enabled VR: Paving the way to cutting the cord," 2018, arXiv:1801.07614. [Online]. Available: https://arxiv.org/abs/1801.07614

[14] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in Proc. International Society for Optics and Photonics, Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, pp. 93–104, 2004.

[15] G. Chaurasia, O. Sorkine, and G. Drettakis, "Silhouette-aware warping for image-based rendering," Comput. Graph. Forum, vol. 30, no. 4, pp. 1223–1232, 2011.

[16] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation,"ACM Trans. Graph., vol. 32, no. 3, p. 30, Jun. 2013.

[17] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 606–619, Mar. 2014.

[18] M. Doma´nski et al., "High efficiency 3D video coding using new tools based on view synthesis," IEEE Trans. Image Process., vol. 22, no. 9, pp. 3517–3527, Sep. 2013.

[19] F. Zou, D. Tian, A. Vetro, H. Sun, O. C. Au, and S. Shimizu, "View synthesis prediction in the 3-D video coding extensions of AVC and HEVC," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 10, pp. 1696–1708, Oct. 2014.

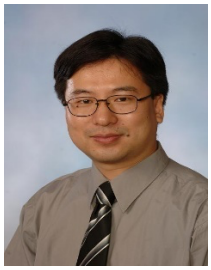
[20] Y. Gao, G. Cheung, T. Maugey, P. Frossard, and J. Liang, "Encoderdriven inpainting strategy in multiview video compression," IEEE Trans. Image Process., vol. 25, no. 1, pp. 134–149, Jan. 2016.

[21] G. Luo, Y. Zhu, Z. Weng, and Z. Li, "A disocclusion inpainting framework for depth-based view synthesis," IEEE Trans. Pattern Anal. Mach. Intell., to be published.

[22] J.-T. Lee, D.-N. Yang, and W. Liao, "Efficient caching for multi-view 3D videos," in Proc. IEEE Global Commun. Conf., Dec. 2016, pp. 1–7.

Jianmei Dai received a B.S. degree in Communication Engineering and an M.S. degree in Communication and Information Systems in 2004 and 2007, respectively, and the Ph.D degree in Information and Communication Engineering from Beijing University of Posts and Telecommunications, Beijing, in 2021. He was a visiting scholar at Auburn University, AL, USA from Apr. 2019 to Apr. 2020. He is currently a lecturer at SEU, Beijing. His research interests include optimization theory and its applications in wireless video transmission and wireless networks.

Zhilong Zhang received the B.E. degree in communication engineering from the University of Science and Technology, Beijing, China, in 2007, and the M.S. and Ph.D. degrees in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, in 2010 and 2016, respectively. From 2010 to 2012, he was a Software Engineer with TD Tech Ltd., Beijing. From 2014 to 2015, he was a Visiting Scholar with Stony Brook University, NY, USA. He is currently a Lecturer with BUPT. His research interests include optimization theory and its applications in wireless video transmission, cross-layer design, and wireless networks.an Editor-at-Large for the IEEE Transactions on Communications.



Shiwen Mao [S' 99-M' 04-SM' 09-F' 19] received his Ph.D. in electrical engineering from Polytechnic University, Brooklyn, NY in 2004. After joining Auburn University, Auburn, AL in 2006, he held the McWane Endowed Professorship from 2012 to 2015 and the Samuel Ginn Endowed Professorship from 2015 to 2020 in the Department of Electrical and Computer Engineering. Currently, he is a professor and Earle C. Williams Eminent Scholar Chair, and Director of the Wireless Engineering Research and Education Center at Auburn University. His research interest includes wireless networks, multimedia communications, and smart grid. He is an Associate Editor-in-Chief of IEEE/CIC China Communications, an Area Editor of IEEE Transactions on Wireless Communications, IEEE Internet of Things Journal, IEEE Open Journal of the Communications Society, and ACM GetMobile, and an Associate Editor of IEEE Transactions on Cognitive Communications and Networking, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Mobile Computing, IEEE Network, IEEE Multimedia, and IEEE Networking Letters, among others. He is a Distinguished Lecturer of IEEE Communications Society and a Distinguished Lecturer of IEEE Council of RFID. He received the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award in 2019 and NSF CAREER Award in 2010. He is a co-recipient of the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the IEEE ComSoc MMTC 2018 Best Journal Award and 2017 Best Conference Paper Award, the Best Demo Award of IEEE SECON 2017, the Best Paper Awards from IEEE GLOBECOM 2019, 2016 & 2015, IEEE WCNC 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.

Danpu Liu received the Ph.D. degree in communication and electrical systems from Beijing University of Posts and Telecommunications, Beijing, China in 1998. She was a visiting scholar at City University of Hong Kong in 2002, University of Manchester in 2005, and Georgia Institute of Technology in 2014. She is currently working at the Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, China. Her research involved MIMO, OFDM as well as broadband wireless access systems. She has published over 100 papers and 3 teaching books, and submitted 26 patent applications. Her recent research interests include 60GHz mmWave communication, wireless high definition video transmission and wireless sensor network.

Translating Video into Language by Enhancing Visual And Language Representations

Pengjie Tang, Yunlan Tan, Jinzhong Li, Bin Tan
College of Electronics & Information Engineering, Jinggangshan University, Ji'an, China
Jiangxi Engineering Laboratory of IoT Technologies for Crop Growth, Ji'an, China
{tangpengjie, tanyunlan, lijinzhong, tanbin}@jgsu.edu.cn

1. Introduction

Video description is to translate given videos into natural language automatically with computer in accordance with visual content. The task is desperate for many fields such as high level video understanding, assistant for impaired human and human–computer interaction, etc. However, it is a real challenging task since the translating depends on not only computer vision but also natural language processing and the resulting complexity of pipeline and framework. In the early days, the template based approaches [1,2] and semantic retrieval based models [3,4] are usually exploited to generate description for videos. But the generation sentences are commonly stereotyped and with poor semantics as the fixed template in advance for the template based models. By contrast, the candidate sentences are more flexible for semantic retrieval based methods where the words, phrases and even sentences are retrieved and recomposed from the beforehand database. But the sentences may be quite different from the real visual content in that it relies on the retrieval database too much.

In this work, the enhanced representation including visual and language method is proposed to improve the quality of generation sentences. Firstly, the pre-trained model is employed to enhance the visual representation sensitive to the general words and sentence patterns, where the CNN model is pre-trained on large scale image classification dataset of Imagenet [5] to prevent the model to be stuck into over fitting and then it is retrained on MSCOCO [6], a dataset for image captioning, based on the model of long recurrent convolutional network (LRCN) [7]. The model pre-trained twice is then employed to extract visual feature for each frame in a video. Secondly, the CNN features of frame sequence are fed into a long short term memory (LSTM) network [8] with double layers for motion feature of the video, in which two architectures including un-factored way and improved factored way are employed to establish the sequential network, and both of the visual and language representations are enhanced to catch not only static and motion visual feature simultaneously but also the personalized linguistic feature. Additionally, a visual sequential mean pooling enhancing method is proposed to seize the possible missing visual details and significant information. Experiments are conducted on MSVD [9] and MSR-VTT [10] datasets, and results demonstrate the effectiveness of the proposed methods with performance improvement compared to the baseline model and other state-of-the-art approaches.

The contributions of this work are concluded as follows. A two-stage pre-training strategy which is more effective for video description is employed. The used CNN model is first pre-trained on Imagenet to prevent the model from falling into over-fitting. And then, the model is fine-tuned on MSCOCO, having the model sensitive to frequently used words and sentence patterns in the task of video captioning. A visual and language representation enhancing method is proposed to capture more comprehensive information including static and motion visual feature and personalized language feature. Additionally, the proposed enhancing method is applied on both un-factored way and factored way LSTM networks for video description. For more visual details and possible missing significant information, a visual sequential mean pooling method is proposed to further improve performance. And a group experiments are implemented on MSVD and MSR-VTT2016 datasets, achieving competitive results compared to other popular state-of-the-art methods.

2. Proposed Model

As presented in Fig. 1. The final visual feature of frames $\langle f_v^1, f_v^2, \dots, f_v^n \rangle$ is then given to bottom LSTM with 1000 hidden outputs of the network. Also, it is concatenated (C1) with the output from the bottom LSTM to enhance the visual representation and capture more static feature and visual detail information. Afterwards, the enhanced representation is transferred to the top LSTM with the same quantity hidden outputs to the bottom layer for further visual feature sequence modeling and refining. When decoding, the same LSTM network is utilized but the outputs from the bottom and top LSTMs at time step $t1$ are sent to the following time step. Before language is fed into the model, the ''one-hot'' method is employed to encode the words coarsely, and the word codes are then fed to an embedding layer (EM) and another fully connected layer (FC2) to extract more abstract language feature and depress language noises. The word feature sequence is subsequently fed to the top LSTM for modeling the sentence structure. The concatenation operation (C2) is conducted again for fusion of output from the top LSTM and word feature at each time step before word predicting

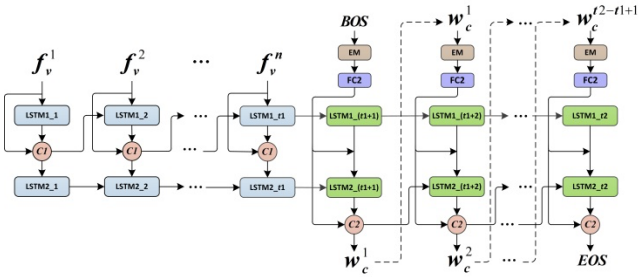


Fig. 1 Enhancing representation model based on un-factored way

As shown in Fig. 2, the $\langle pad \rangle$ representing null (the value is 0 in practice and the dimension is same to output of FC2) is given to the bottom LSTM during visual motion modeling to keep in line with input at decoding stage, while the visual feature sequence $\langle f_v{}^1, f_v{}^2, \ldots, f_v{}^n \rangle$ is sent to the top LSTM and simultaneously concatenated with the just output from the top LSTM to enhance visual representation. When the encoding is over, the output visual representation is just fed into the top LSTM at $t1 + 1$ time step rather than each time step as $h_1{}^{t1}$ in traditional factored way model during decoding. For the word feature sequence $\langle f_w{}^1, f_w{}^2, \ldots, f_w{}^l \rangle$, it is reached to the bottom LSTM firstly and then fused with the just output from the same LSTM. The fusion language representation is following provided to the top LSTM as $x_1{}^{t1+1}$ to prepare to predict words one by one. The architecture not only separates the visual and language representations to a certain extent by covering up the bottom LSTM for visual representation, but also depresses the possible introduced visual noises.

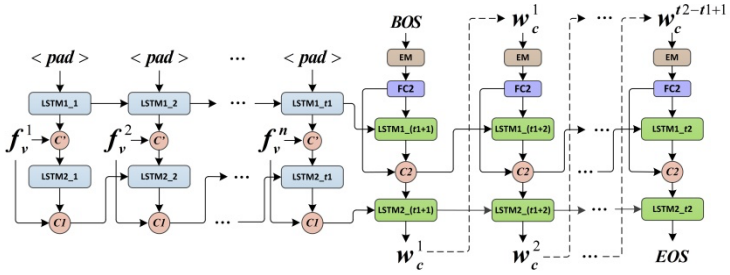


Fig. 2 Enhancing representation model based on improved factored way

The pipeline for visual sequential mean pooling enhancing representation is presented in Fig. 3, where $n1$ is the length of interval for frames, and MP and C0 denote mean pooling operation and feature concatenation operation respectively. The sequence $\langle (mp\_ef)_v{}^1, (mp\_ef)_v{}^2, \ldots, (mp\_ef)_v{}^m \rangle$ is visual representation after mean pooling enhancing, where $m = \lceil n/n1 \rceil$. In detail, the CNN features of all frames in a video are extracted firstly, and the average feature of the frames in the $n1$ interval is then calculated. For each interval, the completion average feature is integrated with CNN feature of the first frame in C0, followed by another fully connected layer (FC1) for dimensionality reduction. The output from FC1 is as the final visual representation and all of them in a video can be formulated as:

$\langle (mp\_ef)_v{}^1, (mp\_ef)_v{}^2, \ldots, (mp\_ef)_v{}^m \rangle = \langle P(f_v{}^1, f_v{}^2, \ldots, f_v{}^{n1}), (f_v{}^{n1+1}, f_v{}^{n1+2}), \ldots, (f_v{}^{(n-1)n1+1}, f_v{}^{(n-1)n1+2}, f_v{}^n) \rangle$

where $(\cdot)$ is for operation of fusion of CNN feature and FC1 transformation. It worth noting that if $mod\,(n, n1) \neq 0$, the length of the last interval is $mod\,(n, n1)$.

3. Experimental Results

The GoogLeNet [11] and ResNet152 [12] models are employed to extract CNN feature of video frames to evaluate the proposed methods under conditions of with visual representation from relative shallow model and deep model respectively. Also, the visual representation enhancing method and language representation enhancing method are implemented alone in models to survey the contributions to the whole system of the two methods. The EFVD-VL stands for the model with both of the two enhancing methods. Analogously, in the model with improved factored way, i-EFVD-VL is used to present the model with the both of enhancing methods respectively. In addition, the proposed visual sequence mean pooling enhancing method is evaluated in this work, where EFVD-VLP is abbreviation of the above mentioned model with un-factored way based on this method.

Comparison with the popular state-of-the-art models on MSVD is shown in Table 1 The performance of proposed methods outperforms most of the other methods even the GoogLeNet feature is employed on multiple evaluation metrics. Particularly, the METEOR comes to 32.3%, exceeding the customary S2VT [13] by 2.5% since the pre-trained CNN model is refined on MSCOCO. It recovers that the optimization on still image

captioning dataset is beneficial to boost sensitive of the later built model for frequently used words and sentence patterns as the generalization of the CNN feature in a video is improved significantly. However, the performance of EFVD-VL with GoogLeNet is not better compared to other methods such as HRNE [14] which performs more excellent on BLEU and METEOR metrics. While when the ResNet152 feature is employed, the proposed models achieve to the new state-of-the-art results on all the evaluation metrics, not least the CIDEr reaches to 78.9% with EFVD-VL, outperforming the GRU-RCN [15] by 11.1%. On the B-4 and METEOR, the EFVD-VL also excels the HRNE [14] by 3.7% and 1.9%. It suggests that more powerful representation is capable of making up for insufficient of used models for sentence generation.

Table 1 Performance (%) comparison with the state-of-the-art methods on MSVD dataset

Method	B-4	METEOR	CIDEr
LSTM-YT [16]	33.3	29.1	--
S2VT [13]	--	29.8	--
HRNE [14]	43.8	33.1	--
GRU-RCN [15]	43.3	31.6	67.8
PickNet [17]	**52.3**	33.3	76.5
EFVD-VL(ResNet152) (our)	**47.5**	**35.0**	**78.9**
i-EFVD-VL(ResNet152) (our)	**47.3**	**35.0**	**77.9**
EFVD-VLP(ResNet152)(our)	**48.3**	**34.5**	**75.9**

The comparison with the other state-of-the-art approaches is given in Table 2. It is evident that the proposed model achieves competitive results, especially the i-EFVD-VL, the performance outperforms most other popular models. However, it is notable that the recent DenseVidCap [18], HRL [19] and HACA [20] possess better performance than the proposed model. Actually, the proposed model is more simple and concise since almost no more extra parameters are introduced into the new model based on S2VT [13]. And the proposed feature enhancing methods can be integrated with models like DenseVidCap [18], and further improve quality of generation sentences.

Table 2 Performance (%) comparison with the state-of-the-art methods on MSR-VTT2016 dataset

Method	B-4	METEOR	CIDEr
MS-RNN [21]	39.8	26.1	40.9
RecNet [22]	39.1	26.6	42.7
DenseVidCap [18]	41.4	28.3	48.9
HRL [19]	41.3	28.7	8.0
HACA [20]	43.4	29.5	49.7
i-EFVD-VL(ResNet152)(our)	40.3	28.4	46.7

4. Conclusion

The motion feature is first extracted by feeding the CNN feature of frame sequence into an LSTM network based on S2VT in this work. Then the visual and language representations are enhanced by fusion of the original feature sequence and output from LSTM to overcome the defect of the static visual information missing and capture the personalized sentence details. Additionally, a visual sequential mean pooling method is also proposed to further enhance the CNN representation for catching more missing visual details and possible significant information. A group of experiments are conducted and the results prove the effectiveness of the proposed methods. It is obvious that the quality of generated sentences no matter with the proposed method alone or with the fusion of the proposed methods is improved greatly compared to the underlying model.

References

[1] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions, Int. J. Comput. Vis. 50 (2) (2002) 171–184.

[2] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, R. Mooney, Integrating language and vision to generate natural language descriptions of videos in the wild, in: Proceedings of International Conference on Computational Linguistics, 2014, pp. 1218–1227.

[3] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele, Translating video content to natural language descriptions, in: IEEE International Conference on Computer Vision, IEEE, 2013, pp. 433–440.

[4] R. Xu, C. Xiong, W. Chen, J.J. Corso, Jointly modeling deep video and compositional text to bridge vision and language in a unified framework, in: Proceedings of Association for the Advancement of Artificial Intelligence, AAAI, 2015, pp. 2346–2352

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[7] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 2625–2634.

[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 09 (08) (1997) 1735–1780.

[9] D.L. Chen, W.B. Dolan, Collecting highly parallel data for paraphrase evaluation, in: The 49th Annual Meeting of the Association for Computational Linguistics, ACL, 2011, pp. 190–200.

[10] J. Xu, T. Mei, T. Yao, Y. Rui, MSR-Vtt: A large video description dataset for bridging video and language, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 5288–5296.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 1–9.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778

[13] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence–video to text, in: IEEE International Conference on Computer Vision, IEEE, 2015, pp. 4534–4542.

[14] P. Pan, Z. Xu, Y. Yang, F. Wu, Y. Zhuang, Hierarchical recurrent neural encoder for video representation with application to captioning, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 1029–1038

[15] N. Ballas, L. Yao, C. Pal, A. Courville, Delving deeper into convolutional networks for learning video representations, in: International Conference on Learning Representations, 2015.

[16] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, in: The 2015 Annual Conference of the North American Chapter of the ACL, ACL, 2015, pp. 1494–1504.

[17] Y. Chen, S. Wang, W. Zhang, Q. Huang, Less is more: Picking informative frames for video captioning, in: European Conference on Computer Vision, Springer, 2018, pp. 367–384.

[18] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, X. Xue, Weakly supervised dense video captioning, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 1916–1924.

[19] X. Wang, W. Chen, J. Wu, Y.-F. Wang, W.Y. Wang, Video captioning via hierarchical reinforcement learning, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 4213–4222.

[20] X. Wang, W. Chen, J. Wu, Y.-F. Wang, W.Y. Wang, Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning, in: Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, 2018, pp. 795–801.

[21] J. Song, Y. Guo, L. Gao, X. Li, H. Alan, H. Shen, From deterministic to generative: Multimodal stochastic RNNs for video captioning, IEEE Trans. Neural Netw. Learn. Syst. 30 (10) (2019) 3047–3058.

[22] B. Wang, L. Ma, W. Zhang, W. Liu, Reconstruction network for video captioning, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 7622–7631.

Pengjie Tang received the M.E. degree in computer software and theory from Nanchang University, Jiangxi, China, in 2009, and the Ph.D. degree in computer science and technology from Tongji University, Shanghai, China, in 2019. He is currently a lecturer of computer science at Jinggangshan University, Ji'an, China. His current research interests include computer vision and visual content understanding.

A Reinforcement-Learning-Based Energy-Efficient Framework for Multi-Task Video Analytics Pipeline

Yingying Zhao [1], Mingzhi Dong[1] , Yujiang Wang[2], Da Feng[3], Qin Lv[4], Robert P. Dick[5], Dongsheng Li[1,6], Tun Lu[1], Ning Gu[1], and Li Shang[1]
[1]Fudan University ,Shanghai,China
[2] Department of Computing, Imperial College London, London, UK.
[3] Alibaba (Beijing) Software Service Company Limited, Beijing, China
[4] University of Colorado Boulder, Boulder,CO, USA
[5] Department of Electrical Engineering and Computer Science College of Engineering, University of Michigan, Ann Arbor, MI, USA
[6] Microsoft Research Asia, Shanghai, China
yingyingzhao@fudan.edu.cn

1. Introduction

Deep learning has achieved great success on video-based computer vision tasks [1], [2], [3]. However, the ability to perform intelligent video analytics in energy-constrained edge devices is becoming increasingly important with the fast expansion of intelligent Internet-of-Things [4], [5]. There is an urgent need for energy-efficient multi-task video analytics.

This work aims to optimize the energy efficiency of video analytics tasks using a variable-resolution strategy. Real-world data redundancy offers us opportunities to optimize energy efficiency via variable-resolution analysis. However, learning appropriate frame resolutions for multi-task video analytics is a challenging problem as appropriate resolutions may vary across different tasks, different scenarios, etc.

In this paper, we propose to use reinforcement learning (RL) to holistically overcome these challenges: (1) complexity variations among different tasks, (2) variable difficulty of different samples, and (3) complicated temporal dynamics. To globally optimize energy efficiency, our RL network learns the best non-myopic policy for determining the spatio-temporal frame resolution of incoming video stream data. Compared with other energy-efficient single-task video analytics solutions [6], [7] that were designed for still images without utilizing temporal information, our work is the first to address the energy consumption optimization problem for multi-task video analytic pipeline, and it is also the first to leverage RL to holistically tackle all these challenges indicated above, and to do end-to-end global efficiency policy optimization.

To evaluate the proposed framework, we have applied it to video instance segmentation [2], a synthesis video analytics pipeline consisting of simultaneous detection, segmentation, and tracking of object instances. Video instance segmentation is considered one of the most challenging multi-task video analytics applications, as it requires the predictions of instance-level segmentation masks while simultaneously tracking and identifying each instance. Our experimental results on the YouTube-VIS dataset [2] indicate that our proposed solution is more energy efficient than all baseline methods.

In summary, this work makes the following contributions:
1)This work presents an adaptive-resolution framework for multi-task video analytics in energy-constrained scenarios. The resulting challenges are managed by Reinforcement Learning (RL) algorithms aiming to globally optimize energy efficiency. To the best of our knowledge, this is the first time that RL has been employed to learn a non-myopic policy for such an energy-efficient framework.
2)We have applied the proposed framework to video instance segmentation [2], one of the most challenging multi-task computer vision tasks. Our framework is significantly more energy efficient than all baseline methods of similar accuracy.

2. Methodology

A.Cumulative Reward

Let $\mathcal{A} = \{a^1, a^2, \dots, a^n\}$ be the set of n potential actions where each action represents using a certain frame resolution, e.g., 1/4 of the original size. We denote the policy of determining frame resolutions as π. Let s_t be the state to be considered by π at time step t, and let $a_t \in \mathcal{A}$ be the decision on the t^{th} frame's resolution, i.e., $a_t = \pi(s_t)$. Let ACC_{a_t} be the performance with a certain metric achieved by decision a_t on that frame, and let E_{a_t} be the energy consumption of that decision. we can define the reward r_t at this time step as

$$r_t = ACC_{a_t} + \lambda \frac{1}{E_{a_t}}, (1)$$

Where λ is a hyper-parameter to trade off accuracy ACC_{a_t} and energy consumption E_{a_t}. A larger r_t is generally more desirable. For a video sequence of length m, our goal is to learn an optimal policy π for maximizing the cumulative rewards G that can be written as

$$G = \sum_{t=1}^{m} \gamma^{t-1} r_t = \sum_{t=1}^{m} \gamma^{t-1} \left(ACC_{a_t} + \lambda \frac{1}{E_{a_t}} \right), (2)$$

Where $\gamma^{t-1} \in [0,1]$ and $a_t = \pi(s_t)$. However, it is difficult to determine a non-myopic policy π for realistic video analytics applications. In this paper, we adopt RL to maximize Equation.

B. Framework Overview

Our goal is to develop an adaptive-resolution multi-task video analytics framework that optimizes energy consumption and accuracy. We model the adaptive resolution selection problem as an MDP. To maximize the cumulative reward G in Equation (2), we use RL to dynamically govern the spatial resolution and temporal dynamics of the complete video instance segmentation pipeline.

Let $\boldsymbol{I} = \{\boldsymbol{I_1}, \boldsymbol{I_2}, \dots, \boldsymbol{I_m}\}$ be a video sequence of length m, where $\boldsymbol{I_t}$ denotes the frame image at time step $t \in Z \cap [1, m]$. For a frame image $\boldsymbol{I_t}$ of resolution $w_t * h_t$, where w_t *and* h_t refer to its width and height, respectively, we define the action set $\mathcal{A} = \{a^1, a^2, \dots, a^k\}$, where a^1 stands for using its original frame size $w_t * h_t$. $a^2, \dots, a^k$ refer to downsampling I_t to a lower resolution, e.g., $\frac{w_t}{2} * \frac{h_t}{2}$, $\frac{w_t}{4} * \frac{h_t}{4}$ and $\frac{w_t}{8} * \frac{h_t}{8}$. We denote the frame where action a^1 is used (i.e., without downsizing the frame image) as the key frame and others as the non-key frames. Therefore, our goal is to find a policy network π_θ that can map the state s_t at each time step t to an appropriate action a_t to maximize the cumulative reward G described in Equation 2. The RL with Double Q-learning (DDQN) [8] is used for optimization. Although various computer vision problems can be solved using this framework, we focus on video instance segmentation.

Specifically, given an incoming frame I_t at time step t, video instance segmentation performs the following prediction tasks: (1)bounding box prediction b_t, (2) object classification c_t, (3) segmentation mask s_t, and (4) tracking prediction d_t. We follow the MaskTrack R-CNN approach [2] to perform these predictions with several modifications. The first step is to use a feature extractor denoted as N_{feat} to extract representative feature descriptors $\boldsymbol{f_t}$, i.e.,$f_t = \mathcal{N}_{feat}(I_t)$. After that, a Regional Proposal Network (RPN) $\mathcal{N}_{\mathcal{RPN}}$ and a RoI Align operation [1] RoIAlign are applied to obtain RoI features f_t' with identical sizes, i.e., $f_t' = RoIAlign\left(\mathcal{N}_{\mathcal{RPN}}(f_t)\right)$. f_t' is then fed into three task-related branches (i.e., heads): (1) the Bounding Boxes Head (BBbox Head) $\mathcal{N}_{bbox}$; (2) the Segmentation Head $\mathcal{N}_{mask}$; and (3) the Tracking Head $\mathcal{N}_{track}$. These three heads generate the required predictions, i.e., $\{b_t, c_t\} = \mathcal{N}_{bbox}(f_t'), s_t = \mathcal{N}_{mask}(f_t')$, and $d_t = \mathcal{N}_{track}(f_t')$. To evaluate the overall performance on frame I_t, we use the metric described by Yang et al. [2]: the mAP score integrating the performance of all four predictions. mAP is higher for more similar bounding boxes.

Following the idea of Deep Feature Flow [9], we also integrate the FlowNet [10] architecture into the MaskTrack R-CNN framework for temporal information inference. Let F be the FlowNet model, and let I_k Ik be the last key frame ($ak = a1$) where the feature descriptor f_k is already computed. If the current frame I_t is determined to be a non-key frame, i.e., $a_t \neq a^1$, we use $\mathcal{F}$ to estimate the optical flow from I_k to I_t denoted as $OF_{k\to t}$ i.e., $OF_{k\to t} = \mathcal{F}(I_t, I_k)$ and the feature descriptor ft is calculated as follows: $f_t = \mathcal{W}(OF_{k\to t}, f_k, S_{k\to t})$, where $\mathcal{W}$ is a warping function and f_t is the scale field from $\boldsymbol{I_k}$ to $\boldsymbol{I_t}$. Zhu et al. [9] give details on the warping function and scale field. If It is determined to be a key frame ($a_k = a^1$), f_t will be obtained from the feature extractor $\mathcal{N}_{feat}$.

Building on the MaskTrackFlow R-CNN, we design a reinforcement-based policy network π_θ with parameters $\theta$to learn appropriate actions at such that the cumulative reward G in Equation 2 can be maximized.

3. Experiments and Results

A. Dataset

We use the YouTube-VIS dataset1 [2] to evaluate the performance of our framework. Since only the training set's annotation is released, we divide the training set with a 90%/5%/5% ratio for training/validation/testing in the following study.

B. Evaluation platform

The proposed framework is designed for energy-constrained edge devices. For evaluation purposes, following Lubana et al. [6], we consider an embedded hardware configuration including a Raspberry Pi 3 equipped with a Sony IMX219 image sensor with variable resolution support. The Sony IMX219 supports a maximum 3,280× 2,464 resolution with 12 MHz clock frequency. As pointed out by Lubana et al. [6], the power consumption in state $P_{sensor,idle}$ is 141.8 mW and that in $P_{sensor,active}$ is 8.27 mW/MP· R+ 17.364 mW + 113.03 mW. We use a T_{exp} of 20 ms in the following study.

The Raspberry Pi 3 is equipped with an embedded GPU consisting of a dedicated image signal processing pipeline [6]. Following prior work [6], we approximate P_{ISP} using the GPU's power consumption. P_{CPU} and P_{GPU} (W-level, typically) can be directly measured by an ammeter. Time required by the Raspberry Pi ISP pipeline is approximately linear in R_{frame} [6], $T_{ISP} = 0.095 \times R_{frame} + 0.032$ (R_{frame} unit is MP).

C. Baselines

We compare the proposed reinforcement-based approach of selecting frame resolutions with the following baseline methods:

(1)Downsampling Scan Method [6]: The Digital Foveation method [6] improves system energy efficiency using a multi-round, variable-resolution, variable-region strategy, in which an application-specific estimated accuracy constraint *(cnstrt)* may be used to govern the sensing and analysis process. It was developed for still images and therefore does not make use of temporal information about downsampling resolution. There are several ways it might be extended to video analytics, and we describe one straight-forward extension for use as a base case. We use the variable-resolution concept of Digital Foveation but gradually vary the sensed resolution for frame I_t from $\frac{w_t}{8} * \frac{h_t}{8}$, $\frac{w_t}{4} * \frac{h_t}{4}$, $\frac{w_t}{2} * \frac{h_t}{2}$ to $w_t * h_t$ if the accuracy reduction has surpassed the constraint *cnstrt*. In this work, we empirically set *cnstrt* to be 0.2, 0.4, 0.6 and 0.8, respectively. We call this extension to video the Downsampling Scan method.

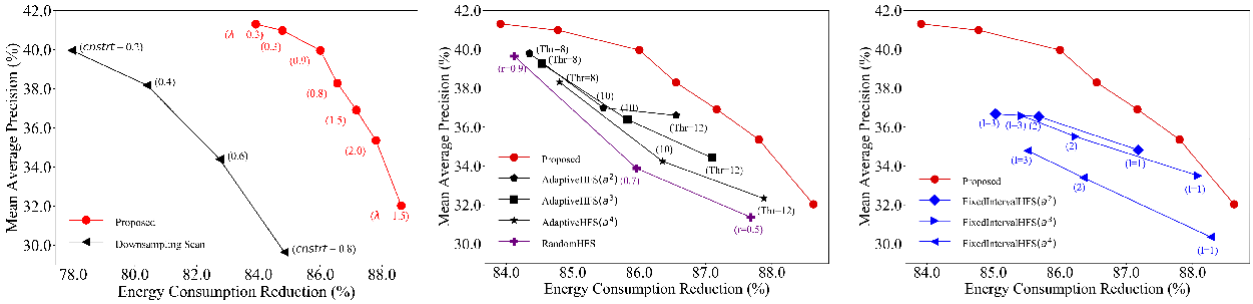
(2)Adaptive High-Resolution Frame Scheduling (AdaptiveHFS): This approach selects the key action a^1 for a frame I_t when the flow magnitude between I_t and the last key frame I_k exceeds a certain threshold *Thr*, otherwise a certain non-key action (i.e., a^2, a^3, or a^4) is taken. Please refer to Xu et al. [11] for the definition of flow magnitude. We select *Thr* from 8 to 12 with an interval of 2. We have three variants of *AdaptiveHFS*, each of which selects a different non-key action to use: *AdaptiveHFS*(a^2), *AdaptiveHFS*(a^3), and *AdaptiveHFS*(a^4).

(3)Fixed-Interval High-Resolution Frame Scheduling(FixIntervalHFS). This baseline method selects a certain non-key action (a^2, a^3, or a^4) for every l ($l \in \{1,2,3\}$) frames, and the rest is set as key action (a^1). According to which non-key action to take, we also have three variants for the *FixIntervalHFS* approach, which are *FixIntervalHFS*(a^2), *FixIntervalHFS*(a^3), and *FixIntervalHFS*(a^4).

(4)Random High-Resolution Frame Scheduling (Ran-omHFS): This baseline method determines actions for each frame randomly with a hybrid distribution. Specifically, for frame I_t, the probability of selecting the key action a^1 is r where $r \in \{0.9,0.7,0.5\}$, and the probability of taking other three non-key actions (a^2, a^3, and a^4) are uniform and sum to $1 - r$.

D.Results

Fig. 1a, Fig. 1b, and Fig. 1c illustrate the mAP (performance) versus energy consumption reduction curves for our method and the baselines. The energy consumption reduces significantly (more than 80%) at the cost of slight accuracy drops, no matter which resolution-selection method is used, thus verifying the effectiveness of the proposed adaptive resolution framework. Note that the policy net only accounts for a very small proportion of the total energy consumption in Fig. 1, which is around 4.2%, thanks to the low-resolution input and its light-weight architecture. Moreover, our method outperforms all other baseline approaches on all the energy consumption intervals, which shows the superiority of our RL-based resolution selector. For realistic computer vision tasks, we can fine-tune the RL models to have different energy consumption rates to suit the varying requirements.



(a)Comparison of our method with the Downsampling Scan method. Different λ values utilized by the proposed method are also annotated

(b) Comparison of our method with the AdaptiveHFS and RandomHFS baselines

(c) Comparison of our method with the FixIntervalHFS baseline

Fig. 1: mAP versus energy consumption reduction between the proposed method and the baselines.

Additionally, as the upper-bound method, MaskTrack R-CNN [2] delivers the highest mAP which is 41.7%. In contrast, our method greatly reduces energy consumption at the cost of slightly reduced accuracy, e.g. our framework achieves 41.4% mAP (only reduced by 0.3%) but saves approximately 84.0% energy consumption at the same time, which is much more energy-efficient.

4. Conclusion

This paper describes an adaptive-resolution energy optimization framework for a multi-task video analytics pipeline in energy-constrained scenarios. We described a reinforcement-learning-based method to govern the operation of the video analytics pipeline by learning the best non-myopic policy for controlling the spatial resolution and temporal dynamics to globally optimize system energy consumption and accuracy. The proposed framework is applied to video instance segmentation which is one of the most challenging video analytics problems. Experimental results demonstrate that our method has better energy efficiency than all baseline methods. This framework can be applied to a wide range of computer vision pipelines with a high demand for efficient energy consumption, e.g., various embedded and Internet-of-Things applications.

References

[1]K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. on Computer Vision, 2017, pp. 2961–2969.
[2]L. Yang, Y. Fan, and N. Xu, “Video instance segmentation,” in Proc. IEEE Int. Conf. on Computer Vision, 2019, pp. 5188–5197.
[3]Q. Wang, C. Yuan, and Y. Liu, “Learning deep conditional neural network for image segmentation,” IEEE Trans. on Multimedia, vol. 21, no. 7, pp. 1839–1852, 2019.
[4]S. Reif, B. Herzog, P. G. Pereira, A. Schmidt, T. Büttner, T. Hönig, W. Schröder-Preikschat, and T. Herfet, “X-Leep: Leveraging cross-layer pacing for energy-efficient edge systems,” in Proc. ACM Int. Conf. on Future Energy Systems. Virtual Event Australia: ACM, Jun. 2020, pp. 548–553. [Online]. Available: https://dl.acm.org/doi/10.1145/3396851. 3402924
[5]R. P. Dick, L. Shang, M. Wolf, and S.-W. Yang, “Embedded intelligence in the Internet-of-Things,” IEEE Design & Test, vol. 37, no. 1, pp. 7–27, 2019.
[6]E. S. Lubana and R. P. Dick, “Digital Foveation: An energy-aware machine vision framework,” IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 37, no. 11, pp. 2371–2380, Nov. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/ 8493507/
[7]E. S. Lubana, V. Aggarwal, and R. P. Dick, “Machine Foveation: An application-aware compressive sensing framework,” in 2019 Data Compression Conf. (DCC). IEEE, 2019, pp. 478–487.
[8]H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” arXiv preprint arXiv:1509.06461, 2015.
[9]X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, “Deep feature flow for video recognition,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2017, pp. 2349–2358.
[10]E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in Proc. IEEE Conf. on Computer Vision and pattern recognition, 2017, pp. 2462–2470.
[11]Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, “Dynamic video segmen-tation network,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 6556–6565.

MMTC OFFICERS (Term 2020 — 2022)

CHAIR

Jun Wu
Fudan University
China

STEERING COMMITTEE CHAIR

Joel J. P. C. Rodrigues
Federal University of Piauí (UFPI)
Brazil

VICE CHAIRS

Shaoen Wu (North America)
Illinois State University
USA

Liang Zhou (Asia)
Nanjing University of Post and Telecommunications
China

Abderrahim Benslimane (Europe)
University of Avignon
France

Qing Yang (Letters & Member Communications)
University of North Texas
USA

SECRETARY

Han Hu
Beijing Institute of Technology
China

STANDARDS LIAISON

Weiyi Zhang
AT&T Research
USA

MMTC Communication-Frontier BOARD MEMBERS (Term 2016—2018)

Danda Rawat	Director	Howard University	USA
Sudip Misra	Co-Director	IIT Kharagpur	India
Guanyu Gao	Co-Director	Nanjing University of Science and Technology	China
Rui Wang	Co-Director	Tongji University	China
Lei Chen	Editor	Georgia Southern University	USA
Tasos Dagiuklas	Editor	London South Bank University	UK
ShuaiShuai Guo	Editor	King Abdullah University of Science and Technology	Saudi Arabia
Kejie Lu	Editor	University of Puerto Rico at Mayagüez	Puerto Rico
Nathalie Mitton	Editor	Inria Lille-Nord Europe	France
Zheng Chang	Editor	University of Jyväskylä	Finland
Dapeng Wu	Editor	Chongqing University of Posts & Telecommunications	China
Luca Foschini	Editor	University of Bologna	Italy
Mohamed Faten Zhani	Editor	University of Quebec	Canada
Armir Bujari	Editor	University of Padua	Italy
Kuan Zhang	Editor	University of Nebraska-Lincoln	USA
Bin Tan	Editor	Jinggangshan University	China