MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://mmc.committees.comsoc.org/

MMTC Communications – Review

Vol. 11, No. 2, April 2020

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the April 2020 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises a number of reviews that cover multiple facets of multimedia communication research including beamforming, AI, video coding, etc. These reviews are briefly introduced below.

The first paper, published in IEEE Transactions on Wireless Communications and edited by Dr. Yiming Liu and Dr. Rui Wang, developed an algorithm to optimize the active transmit beamforming at the AP and passive reflect beamforming by the phase shifters at the IRS to minimize the total transmit power.

The second paper is published in IEEE Transactions on Multimedia and edited by Dr. Mukesh Saini. It proposes a novel approach for auto-curating sports highlights and demonstrates with creating a real-world system for the editorial aid of golf and tennis highlight reels.

The third paper, published in IEEE Transactions on Multimedia and edited by Dr. Carl James Debono, investigates a solution to reduce the coding complexity of inter prediction for SHVC configured for quality scalability.

The fourth paper, published in CVPR conference and edited by Dr. Jun Zhou, points out that a large and complex network can be split into smaller and simpler networks in order to bring the benefit of low computational cost and improved performance.

The fifth paper, published in IEEE Transactions on Multimedia and edited by Dr. Bruno Macchiavello, investigates the use of a single kernel on convolution LSTM networks solving the problem of the choice of the optimal kernel size.

The sixth paper, published in IEEE IoTs Journal and edited by Dr. Jinbo Xiong, studies the selective sensing and its unique advantages while attracting wide attention.

The seventh paper, published in ACM Transactions on Graphics and edited by Dr. Carsten Griwodz, explores the symmetric registration technique of iterative closed points and the ability to work with point clouds.

All the authors, nominators, reviewers, editors, and others who contribute to the release of this issue deserve appreciation with thanks.

IEEE ComSoc MMTC Communications – Review Directors

Qing Yang
University of North Texas, USA
Email: qing.yang@unt.edu

Roger Zimmermann
National University of Singapore, Singapore
Email: rogerz@comp.nus.edu.sg

Wei Wang
San Diego State University, USA
Email: wwang@mail.sdsu.edu

Zhou Su
Shanghai University, China
Email: zhousu@ieee.org

Transmit and Reflect Beamforming Design for Intelligent Reflecting Surface

A short review for "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming" (Edited by Yiming Liu and Rui Wang)

Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," IEEE Transactions on Wireless Communications, vol. 18, no. 11, pp. 5394-5409, Nov. 2019.

To satisfy the requirements of enhanced mobile broadband, ultra-reliability and low-latency, and massive machine type communications in the forthcoming 5th generation (5G) networks, a large number of wireless communication technologies have been proposed and thoroughly investigated in the last decade, including most prominently the ultra-dense network (UDN), massive multiple-input multiple-output (MIMO), and millimeter wave (mmWave) communication [1]. However, the requirements of high hardware cost and high complexity are still the main hindrances to the implementations [2], which necessitates radically new communication paradigms, especially at the physical layer.

The recent advances in radio frequency micro electromechanical systems and metamaterial have made the reconfigurability of reflecting surfaces possible, even by controlling the phase shifters in real time [3]. Thus, intelligent reflecting surface (IRS) can be used as a promising new solution to improve the energy and spectrum efficiency with less complexity and hardware cost. Specifically, IRS is a planar array consisting of many reconfigurable passive elements, *e.g.*, low-cost printed dipoles, where each of the elements can induce a certain phase shift independently on the incident signal, thus collaboratively changing the reflected signal propagation [4]. However, the research on the performance analysis of IRS-assisted wireless communication systems and the optimization for transmit and reflect beamforming design is still in its infancy.

In this paper, an IRS-assisted multi-user multiple-input single-output (MISO) system is considered in a single cell, where a multi-antenna access point (AP) serves multiple single-antenna users with the help of an IRS. Each user receives the superposed signals from the direct link and the reflected link (AP-user link and AP-IRS-user link). The authors jointly optimize the active transmit beamforming at the AP and passive reflect beamforming by the phase shifters at the IRS to minimize the total transmit power at the AP with a tolerable signal-to-interference-plus-noise ratio (SINR) at the user receivers. One major contribution of this paper is to analyze the transmit and reflect beamforming design in two special cases: If the channel of the direct link is much stronger than that of the reflected link, the AP should beam toward the user directly; If the direct link is blocked by obstacles, the AP ought to adjust its beam toward the IRS, and the IRS should focus the signal into a sharp beam toward the user to achieve a high beamforming gain.

While in the case with more general setup, the joint optimization of the transmit beamforming at the AP and reflect beamforming at the IRS is difficult, due to the non-convex SINR constraints as well as the signal unit-modulus constraints imposed by passive phase shifters. Although the existing studies on constant-envelope precoding [5], [6] and hybrid digital/analog processing [7], [8] have studied the beamforming optimization under unit-modulus constraints, such designs are mainly restricted to either the transmitter or the receiver side. The joint active and passive beamforming optimization at both the AP and IRS has not been addressed.

The second contribution of this paper is to address the above problem in the case of single-user. The authors apply the semidefinite relaxation (SDR) technique to obtain a high-quality approximate solution as well as a lower

bound of the optimal value to evaluate the tightness of approximate solutions. To reduce the computational complexity, the authors further propose an efficient algorithm based on the alternating optimization of the phase shifts and transmit beamforming vector in an iterative manner, where their optimal solutions are derived in closed-form with the other being fixed.

The third contribution of this paper is to extend the transmit and reflect beamforming design for the single-user case to the multi-user case. To obtain suboptimal solutions, the authors propose two algorithms: *Alternating Optimization Algorithm* and *Two-Stage Algorithm*, which offer different tradeoffs between performance and complexity. By using the *Alternating Optimization Algorithm*, the transmit beamforming direction and transmit power at the AP are optimized iteratively with the phase shifts at the IRS in an alternating manner, until the convergence is achieved. By using the *Two-Stage Algorithm*, the joint optimization problem of the beamforming design is divided into two beamforming subproblems, for optimizing the transmit beamforming and reflect beamforming, respectively. Compared to the *Alternating Optimization Algorithm*, the *Two-Stage Algorithm* has lower computational complexity, but may suffer from certain performance loss.

The authors finally validate the proposed solutions, and their numerical results demonstrate that the required transmit power at the AP to meet users' SINR targets can be considerably reduced by deploying the IRS as compared to the conventional setup without using IRS for both single-user and multi-user setups. The authors' results also show that the AP's transmit power decreases with the number of reflecting elements N at the IRS in the order of N^2 when N is sufficiently large, which is also consistent with the performance scaling law derived analytically by the authors.

References:

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74-80, February 2014.

[2] S. Zhang, Q. Wu, S. Xu and G. Y. Li, "Fundamental Green Tradeoffs: Progresses, Challenges, and Impacts on 5G Networks," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 33-56, Firstquarter 2017.

[3] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light: Science & Applications*, vol. 3, no. 10, p. e218, October 2014.

[4] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini and R. Zhang, "Wireless Communications Through Reconfigurable Intelligent Surfaces," *IEEE Access*, vol. 7, pp. 116753-116773, August 2019.

[5] S. K. Mohammed and E. G. Larsson, "Single-User Beamforming in Large-Scale MISO Systems with Per-Antenna Constant-Envelope Constraints: The Doughnut Channel," *IEEE Transactions on Wireless Communications*, vol. 11, no. 11, pp. 3992-4005, November 2012.

[6] S. Zhang, R. Zhang and T. J. Lim, "Constant Envelope Precoding for MIMO Systems," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 149-162, January 2018.

[7] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi and R. W. Heath, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499-1513, March 2014.

[8] F. Sohrabi and W. Yu, "Hybrid Digital and Analog Beamforming Design for Large-Scale Antenna Arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501-513, April 2016.

Yiming Liu, received the B.E. degree in communication engineering from Hohai University, China, in 2019. He is currently with the College of Electronics and Information Engineering, Tongji University, China, and studying in Artificial Intelligence and Blockchain Intelligence Labs, and Positioning and Navigation Labs. His research interests include information theory and signal processing, advanced equalization, channel estimation and synchronization techniques, artificial intelligence and

machine learning for wireless communication systems, blockchain technology, and semantic access control and authentication.

Rui Wang received his Ph.D. degree in 2013 from Shanghai Jiao Tong University, China. From Aug. 2012 to Feb. 2013, he was a visiting Ph.D. student at the Department of Electrical Engineering of University of California, Riverside. From Oct. 2013 to Oct. 2014, he was with the Institute of Network Coding, the Chinese University of Hong Kong as a post- doctoral research associate. From Oct. 2014 to Dec. 2016, he was with the College of Electronics and Information Engineering, Tongji University as an assistant professor, where he is currently an associate professor.

Using Multimodal Artificial Intelligence to Generate Highlights of Sport Events

A short review for "Automatic Curation of Sports Highlights using Multimodal Excitement Features"

Edited by Mukesh Saini

Michele Merler, Khoi-Nguyen C. Mac, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, Jinjun Xiong, Minh N. Do, John R. Smith and Rogerio S. Feris, "Automatic Curation of Sports Highlights using Multimodal Excitement Features", IEEE Transactions on Multimedia, vol 21, issue 5, pages 1147-1160, October 2018.

The tremendous growth of video data has resulted in a significant demand for tools that can accelerate and simplify the production of sports highlight packages for more effective browsing, searching, and content summarization. In a major professional golf tournament such as Masters, for example, with 90 golfers playing multiple rounds over four days, video from every tee, every hole and multiple camera angles can quickly add up to hundreds of hours of footage. The production of sports highlight packages summarizing a game's most exciting moments is labor-intensive video editing.

The authors propose a novel approach for auto-curating sports highlights and demonstrate it to create a real-world system for the editorial aid of golf and tennis highlight reels. The proposed method fuses information from the players' reactions (action recognition such as high-fives and fist pumps), players' expressions (aggressive, tense, smiling, and neutral), spectators (crowd cheering), commentator (tone of the voice and word analysis), and game analytics to determine the most interesting moments of a game. The system identifies the start and end frames of key shot highlights with additional metadata. The additional metadata could be in the form of a player's name or hole number in the video frame. Alternatively, the metadata is obtained from match analysts and statisticians. In addition, the authors exploit multiple modalities for training various classifiers with reduced labeled samples.

The approach introduced by the authors combines information from the player, spectators, and the commentator to determine a game's most exciting moments. The excitement level of video segments is based on the weighted combination of excitement scores produced by deep learning classifiers built on top multimodal audio and visual markers. For audio markers, crowd cheering is perhaps the most veritable form of approval of a player's shot within the context of any sport. Another important audio marker is excitement in the commentators' tone while describing a shot. The authors leverage SoundNet [1] to construct audio-based classifiers for crowd and commentator excitement, trained on a dataset appositely constructed for this work. For visual markers, the authors model player reaction and facial expressions. Visual action recognition of player's celebration (such as high fives or fist pumps) is detected in individual frames. Classifiers to detect player's celebration are based upon the VGG-16 and the ResNet-50 architectures pretrained on ImageNet and fine-tuned on a dataset annotated specifically for celebrations. Facial expression recognition is trained using player faces extracted from the action celebration images. The facial expression classifier was trained by fine-tuning a VGG-face [2] as aggressive, tense, smiling, and neutral.

While the audio and visual classifiers were independent modules within the system, the training data gathering process proceeded by several rounds of bootstrapping, which exploited the correlation among modalities. In particular the authors applied this principle for the visual player celebration classifier training. The underlying assumption is that the likelihood of finding frames showing players celebrating is higher in frames close to a detected crowd cheer. Therefore, in each bootstrapping round, the authors fed unlabeled videos to the current classifiers. The examples with highest positive or negative scores for each individual classifier are sent for annotation. The newly labeled data was then used to finetune the classifiers for the next bootstrapping round, until a certain accuracy level on a held-out validation set was reached.

Additional metadata was gathered through speech-to-text from commentaries, automatically recognizing exciting words or expressions used, (such as "beautiful shot"), game analytics and text OCR. Finally, the system enables personalized highlight retrieval or alerts based on player name, hole or field number, location, and time. It was successfully demonstrated and deployed at major international golf and tennis tournaments since 2017, namely the Golf Masters, Wimbledon, and the US Open. As an example, for the 2017 Masters, the system analyzed in near real-time the content of four channels broadcasting simultaneously over the course of four consecutive days, for a total of 124 hours of content, and automatically produced 741 highlights over all channels and days.

The author's compare the clips automatically generated by the system with the collection of highlights professionally produced by the official Masters curators and published on their Twitter channel through human evaluation and ranking. For the 2017 Masters, 54% of the clips selected by the system overlapped with the official highlights reels. Furthermore, user studies showed that 90% of the non-overlapping ones were of the same quality as the official highlights. Similarly, the automatic selection of clips for highlights of 2017 Wimbledon and 2017 US Open agreed with human preferences 80% and 84.2% of the times, respectively.

It should be noted that both tennis and golf are relatively quiet sports, where exciting events are rare. A sport like basketball or soccer has the crowds chanting all the time and it would be challenging to directly employ a completely sport-agnostic system like the proposed without any adaptation. In those instances, specialized knowledge of the sport in question can definitely add value to the highlight selection process. An integration of sport- specific action detection markers (i.e. a basketball dunk, or a soccer goal) might be helpful to extend the system.

Acknowledgement:

The R-Letter Editorial Board thanks Mukesh Saini for nominating this work.

References:

[1] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in NIPS, 2016.
[2] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in BMVC, 2015.

Mukesh Saini is an Assistant Professor at IIT Ropar. His research area is multimedia systems, with specialization in surveillance, automatic movie making, and smart classrooms. Mukesh Saini obtained his Master of Technology (M. Tech) in Electronics Design and Technology from Indian Institute of Science (IISc), Bangalore, in 2006 and PhD in Computer Science from School of Computing, National University of Singapore, Singapore in 2012. Mukesh has been actively serving as Reviewer, TPC member, Tutorial organizer, and Panelist for various reputed conferences and journals.

Reducing Complexity in Scalable Video Coding

A short review for "Fast Depth and Inter Mode Prediction for Quality Scalable High Efficiency Video Coding"

Edited by Carl James Debono

D. Wang, Y. Sun, C. Zhu, W. Li and F. Dufaux, "Fast Depth and Inter Mode Prediction for Quality Scalable High Efficiency Video Coding," IEEE Transactions on Multimedia, vol. 22, no. 4, pp. 833 – 845, April 2020.

The High Efficiency Video Coding (HEVC) standard allows for efficient encoding of high definition and ultra-high definition video content. This efficiency comes at the expense of increased complexity [1]. One of the extensions of the HEVC standard is the Scalable High Efficiency Video Coding (SHVC) that uses multiple layers and inter-layer prediction in its coding strategy. This further increases the complexity as more computations need to be done to determine the most efficient way to encode the video content while maintaining the quality. Reducing the complexity of SHVC is therefore important to make it more appealing to use in applications such as video broadcasting, multiuser video conferencing and video streaming.

Transmission of video to many users means that the content must be consumed on a number of devices, each having different characteristics. The recipient devices have different processing resources and screen resolutions. Moreover, the data passes on different networks offering different data rates and quality of service. Therefore, to satisfy the needs of all users, different video streams encoded with different qualities and resolutions need to be sent requiring a large amount of bandwidth, storage and processing. Alternatively, we can use Scalable Video Coding (SVC) [2] to encode a base layer with enhancement layers and the receiving device decodes the layers that it needs from the video stream. SVC supports spatial scalability with the base layer providing a low resolution and the enhancement layers progressively increasing this resolution, temporal scalability with the base layer providing low frame rates that progressively improve with the enhancement layers and quality scalability with the base layer offering the lowest quality. SHVC supports the same scalability function of SVC but adds bit-depth scalability and color gamut scalability [3].

Some research work tried to reduce the complexity of SHVC. The work in [4] uses the rate-distortion information from the neighboring blocks of the current block being encoded in the enhancement layer together with the corresponding block at the base layer and its four neighbors to predict its rate-distortion cost. In [5] the modes are used for prediction instead of the rate-distortion metric. Spatial and inter-layer correlations are used in [6] to predict the quad-tree structure of the coding tree units. Other techniques use probabilistic models [7 – 9] to reduce the complexity of the SHVC.

The authors of the original paper propose a solution to reduce the coding complexity of inter prediction for SHVC configured for quality scalability. The main contributions are summarized as (1) combining the inter-layer correlation with spatial correlation and its correlation degree for prediction; (2) depth correlation and the distribution of the residual are used to predict inter-layer reference (ILR) and merge modes; (3) square partitions are terminated early depending on the rate-distortion cost differences; (4) non-square partitions are deduced from the difference in expected values of the residual coefficients of the square partitions; and (5) depth early termination is predicted from inter-layer and spatial correlations combined with the distribution of the residual.

The algorithm proposed by the authors starts by predicting the depth candidates using the correlation-based depth prediction strategy. The selected depth candidates are used to check whether the best mode is in ILR and merge modes. Other modes are skipped if this is true, if

not the inter2N×2N mode is tested and the algorithm checks whether the best modes lies in the square modes. If this is true the non-square modes are not tested while if negative these are checked. After checking the depth, the residuals of the likelihood depth predicted by correlation are checked to establish whether this is the best depth for early termination.

The authors report results that show that the solution manages to achieve coding times that are on average 71.14% and 67.43% lower when tested on two different sets of quantization parameters. This comes at a small loss of coding efficiency. The experiments were conducted using the common SHM test conditions.

Further improvement is video compression complexity is needed for SHVC to make it feasible in real-time bandwidth-limited applications. This can be achieved by using more advanced machine learning solutions and using hybrid techniques.

References:

[1] G. Correa, P. Assuncao, L. Agostini, and L. S. Cruz, "Performance and computational complexity assessment of high efficiency video encoders," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1899–1909, December 2012.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, September 2007.

[3] J. M. Boyce, Y. Ye, and J. L. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable extensions of the high efficiency video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34, January 2016.

[4] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos, "Content adaptive complexity reduction scheme for quality/fidelity scalable HEVC," *document JCTVC-L0042, ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG11*, January 2013.

[5] H. R. Tohidypour, H. Bashashati, M. T. Pourazad, and P. Nasiopoulos, "Fastmode assignment for quality scalable extension of the high efficiency video coding (HEVC) standard: A Bayesian approach," in *Proceedings of the 6th Balkan Conference in Informatics*, September 2013, pp. 61–65.

[6] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos, "Probabilistic approach for predicting the size of coding units in the quad-tree structure of the quality and spatial scalable HEVC," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 182–195, February 2016.

[7] H. R. Tohidypour, H. Bashashati, M. T. Pourazad, and P. Nasiopoulos, "Online-learning-based mode prediction method for quality scalable extension of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 2204–2215, October 2017.

[8] D. Wang, C. Zhu, Y. Sun, F. Dufaux, and Y. Huang, "Efficient multistrategy intra prediction for quality scalable high efficiency video coding," IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 2063–2074, April 2019.

[9] C.-H. Yeh, W.-Y. Tseng, L.-W. Kang, C.-W. Lee, K. Muchtar, and M.-J. Chen, "Coding unit complexity-based predictions of coding unit depth and prediction unit mode for efficient HEVC-to-SHVC transcoding with quality scalability," *Journal of Visual Communication and Image Repersentation*, vol. 55, pp. 342–351, August 2018.

Carl James Debono (S'97, M'01, SM'07) received his B.Eng. (Hons.) degree in Electrical Engineering from the University of Malta, Malta, in 1997 and the Ph.D. degree in Electronics and Computer Engineering from the University of Pavia, Italy, in 2000.

Between 1997 and 2001 he was employed as a Research Engineer in the area of Integrated Circuit Design with the Department of Microelectronics at the University of Malta. In 2000 he was also engaged as a Research Associate with Texas A&M University, Texas, USA. In 2001 he was appointed Lecturer with the Department of Communications and Computer

Engineering at the University of Malta and is now a Professor. He is currently the Dean of the Faculty of ICT at the University of Malta.

Prof. Debono is a senior member of the IEEE and served as chair of the IEEE Malta Section between 2007 and 2010. He was the IEEE Region 8 Vice-Chair of Technical Activities between 2013 and 2014. He has served on various technical program committees of international conferences and as a reviewer in journals and conferences. His research interests are in multi-view video coding, resilient multimedia transmission, computer vision, and modeling of communication systems.

Distinguishing Focused and Blurred Regions in an Image

A short review for "Enhancing Diversity of Defocus Blur Detectors via Cross-Ensemble Network"

Edited by Jun Zhou

Wenda Zhao, Bowen Zheng, Qiuhua Lin, and Huchuan Lu, "Enhancing Diversity of Defocus Blur Detectors via Cross-Ensemble Network," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8897-8905, Long Beach, CA, USA, 2019.

Focusing is a fundamental function of cameras. In many photo taking practices, the focus of camera is on a specific area in the scene. As a consequence, other areas in the captured image are out of focus, causing image blur. Varied extent of blurriness creates the depth effect in an image, which is an essential aspect of the art of photography.

From the computer vision point of view, the information of focus is very important for many applications, for example, image quality assessment [1], depth estimation [2] and object detection [3]. Effective usage of the focus information requires estimation of defocus blur throughout the image or detection of regions that are focused. Such regions normally correspond to one or more objects with clear semantic meaning.

So far quite a few methods have been proposed to tackle the defocus blur detection problem with different considerations. Like solutions to other computer vision problems, they can be divided into hand-crafted feature based or deep learning based. For hand-crafted feature based methods, gradient, frequency, and texture features are often used as they are closely related to edges from which blurriness can be estimated with relatively higher accuracy [4]. Deep learning based approaches directly learn image representation and perform feature extraction without the need to design features. They can be enriched with various mechanisms, e.g. attention modelling [1] and multi-scale analysis [4], to achieve improved detection results.

Despite the advantage of deep learning models, they normally require deep and wide network structure [5], which causes high computational complexity and lack of diversity in handling different types of blurs. This motivates the development of the defocus blur detector cross-ensemble network (DBD-CENet) proposed in this reviewed paper by Zhao *et al*. The overall idea of DBD-CENET is that a large network can be broken down into many smaller networks, each correspond to a different defocus blur detector. These detectors not only introduce diversity to the blur detection, but also detection errors. When these detectors are combined, the final detection accuracy can be improved, and at the same time, the computational cost can be reduced.

The design of DBD-CENet adopts an end-to-end network with two key components. The first is a feature extraction network which extracts low-level image features. These features are fed into the second component, which contains two parallel subnetworks, for defocus blur estimation. Each subnetwork contains a group of defocus blur detectors, and each detector makes an estimation. The training of the detector is based on a cross-negative loss and a self-negative correlation loss. An error function is introduced to penalize the correlation of each detector with the other detectors in the same and different subnetworks, so that the diversity of estimation can be enhanced.

In the model implementation, a modified VGG16 network is used as the baseline. The first two convolutional blocks are used for low-level feature extraction. The third to the fifth layers are split into two subnetworks used for blur detection. In each subnetwork, a sixth layer is used to generate different detectors. The authors implemented three models, i.e., single detector network, multi-detector ensemble network, and cross-ensemble network. Single detection is implemented as a large single network. The multi-detector ensemble network contains multiple detectors for blur estimation in one network. Finally, the cross-ensemble network consists of dual subnetworks.

The proposed method was tested on DUT defocus blur dataset [5] and CUHK blur dataset

[6]. Comparison with six state-of-the-arts defocus detection methods show that the method in this paper has achieved the best results in terms of both quantitative measures and qualitative evaluation. The method demonstrated particular advantages in handling smooth regions and multi-scale objects.

This paper showcases a successful work that utilizes the idea of ensemble learning. With deep neural networks being deployed in hardware platforms with limited computational resources, reducing the complexity of network becomes an important task. This paper points out that a large and complex network can be split into smaller and simpler networks in order to bring the benefit of low computational cost and improved performance.

References:

[1] S. Zhang, X. Shen, Z. Lin, R. Měch, J. Costeira, and J. Moura. "Learning to understand image blur", IEEE Conference on Computer Vision and Pattern Recognition, pp. 6586-6595, 2018.

[2] A. Zia, J. Zhou, and Y. Gao. "Relative depth estimation from hyperspectral data", International Conference on Digital Image Computing: Techniques and Applications, pages 1-7, 2015.

[3] C. Tang, P. Wang, C. Zhang, and W. Li. "Salient object detection via weighted low rank matrix recovery", IEEE Signal Processing Letters, Vol. 24, No. 4, pp. 490–494, 2017.

[4] C. Tang, X. Zhu, X. Liu, L. Wang and A. Zomaya, "DeFusionNET: defocus blur detection via recurrently fusing and refining multi-scale deep features", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2695-2704, 2019.

[5] W. Zhao, F. Zhao, D. Wang, and H. Lu. "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network". IEEE Conference on Computer Vision and Pattern Recognition, pp. 3080–3088, 2018.

[6] J. Shi, L. Xu, and J. Jia. "Discriminative blur detection features", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2965–2972, 2014.

Jun Zhou received the B.S. degree in computer science and the B.E. degree in international business from Nanjing University of Science and Technology, China, in 1996 and 1998, respectively. He received the M.S. degree in computer science from Concordia University, Canada, in 2002, and the Ph.D. degree in computing science from University of Alberta, Canada, in 2006.

He is now an associate professor in the School of Information and Communication Technology in Griffith University. Prior to this appointment, he had been a research fellow in the Australian National University, and a researcher at NICTA, Australia. His research interests are in spectral imaging, pattern recognition, computer vision, and their applications to and environmental informatics and remote sensing.

A Multi-Kernel Approach for Convolutional LSTM Networks Applied to Action Recognition

A short review for "Deep Multi-Kernel Convolutional LSTM Networks and an Attention-Based Mechanism for Videos"

Edited by Bruno Macchiavello

S. Agethen and W. H. Hsu, "Deep Multi-Kernel Convolutional LSTM Networks and an Attention-Based Mechanism for Videos," in IEEE Transactions on Multimedia, vol. 22, no. 3, pp. 819-829, March 2020.

Nowadays, video data is more frequently encountered in our everyday lives, due to social media platforms and video streaming services. Therefore, automatic video processing is rapidly becoming a necessity. One highly active topic on automatic video processing is action recognition. Many applications can benefit from action recognition such as: intelligent video surveillance, autonomous driving, sport analysis and human-computer interaction.

The fundament of human action recognition is capturing the spatial body features and its temporal evolution in a video [1]. To capture the information in videos, handcrafted global features such as the histogram of oriented gradients [2] can be used. However, the presence of noise can significantly affect the performance of this technique. In order to mitigate this issue, the use of local features has been investigated [3 – 4]. More modern methods rely on deep features [5 – 7]. Two approaches are more common: deep convolutional networks (DCNs) and recurrent networks with short-term memory (LSTM networks). Several variations of LSTM cells exist. Recently, convolutional LSTM cells have been proposed [8]. These cells add the benefits of convolution to a LSTM unit. The authors state that to the best of their knowledge, previous works based on convolutional LSTM cells for action recognition have always used a single kernel. In this work the authors propose a multi-kernel approach. In order to support the multi-kernel architecture, they introduce an additional depth for both input-to-hidden and hidden-to-hidden processing and a kernel attention mask.

The authors first investigate the use of a single kernel on convolution LSTM networks. The main problem is the choice of the optimal kernel size. A kernel size that is too large results in the system degenerating into the original fully connected formulation. If the kernel is too small, the kernel will not be able to capture all the information. The authors argue that the exclusive use of kernel of one particular size is not optimal for convolutional LSTM. Instead, they suggest using an array of kernels of different sizes. They state that if we consider a video showing two objects, where the objects are moving at significantly different speeds. A small kernel is unable to link a fast-moving object during the transition from timestep t to timestep $t + 1$, arguably because it has moved too far. However, always using large kernels may be disadvantageous because they require more parameters, are significantly slower and can degenerate into fully connected layers. In order to support their hypothesis, the authors constructed a synthetic dataset of sequences of moving digits. A future predictor is created in form of an encoder-decoder architecture using different kernel sizes. The associated loss for smaller kernel sizes grew significantly faster than for larger kernel sizes. Concurrently, the computational costs were much higher for the larger kernels. While, this experiment does not prove the hypothesis, but it does show that it has merit.

The authors propose a multi-kernel configuration. They used multiple kernels in parallel and concatenate the individual results. However, a naive concatenation is certainly not optimal. The authors propose two strategies: *(i)* interleave the result by first splitting each kernel´s output and then concatenation in the correct order; and *(ii)* use a 1×1 convolution operation to integrate the individual results.

One other problem to the multi-kernel approach is that any kernel learns on all regions of an

image, even if it is not optimally suited for this image region. The authors want to enforce large kernels to concentrate on faster objects and smaller kernels to concentrate on slower objects. To determine the utility of a kernel on a particular image region, the authors generate attention masks from optical flow features. The magnitude of the optical flow determines the distance a particular pixel has moved on the *x*- or *y*-axis, and therefore
represents speed. To generate each mask, they employ a DCN on the optical flow features. This mask can then be applied by elementwise multiplication. Their goal is that through backpropagation of error, each mask specializes onto the corresponding convolutional kernel.

For evaluation, the authors use to different datasets. However due to the small size of the data set they started from pre-trained models and use the training datasets for refinement. They use their multi-kernel approach in two previously proposed architectures, namely the VGG-16 [9] and the I3D [10]. For the VGG-based configuration the proposed multi-kernel action recognition approach outperforms a single kernel by up to *2.82%* (depending on the data set and kernel size). For the I3D-based configuration using an ento-to-end training they were able to outperform the baseline configuration by *1.69%*. The computational complexity is also analyzed. As expected, the processing time for mixing kernels of two sizes is between the required times for the traditional convolutional LSTM layers of the respective kernel sizes. The overhead is minimal and in the order of *1%*.

In conclusion in this work the authors analyzed the problem of different speeds in convolutional LSTM and proposed replacing the single convolutional kernel with a set of kernels of different sizes. They also presented an attention-based method that is specifically tailored to their system. Finally, future studies may need to solidify the hypothesis in this work, and also address

References:

[1] Z. Zhang et al., "Deep learning based human action recognition: A survey," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 3780-3785.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Comput. Soc.Conf.Comput. Vis. Pattern Recognit., Jun. 2005, vol. 1, pp. 886–893.

[3] A. Klaser, M. Marszalek, and C. Schmid, “A spatio-temporal descriptor based on 3D-gradients,” in Proc. BMVC - 19th Brit. Mach. Vis. Conf., Sep. 2008, pp. 275:1–10.

[4] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 1948–1955.

[5] J. Y.-H. Ng et al., “Beyond short snippets: Deep networks for video classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4694–4702.

[6] A. Karpathy et al., “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1725–1732.

[7] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., 2014, vol. 1, pp. 568–576.

[8] X. Shi et al., “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, vol. 1, pp. 802–810.

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

[10] J. Carreira and A. Zisserman, “Quo vadis, action recognition?Anewmodel and the kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 4724–4733.

Bruno Macchiavello, is an associate professor at the Department of Computer Science of the University of Brasilia (UnB), Brazil. He received his B. Eng. degree in the Pontifical Catholic University of Peru in 2001, and the M. Sc. and D.Sc. degrees in electrical engineering from the University of Brasilia in 2004 and 2009, respectively. Prior to his current position he helped develop a database system for the Ministry of Transport and Communications in Peru. He is and Are Editor for the Elsevier Journal Signal Processing: Image Communications. His main research interests include video and image coding, image segmentation, distributed video and source coding, multi-view and 3D video processing. He is currently head of the Graduate Program of Informatics at UnB.

An Attention Mechanism Inspired Selective Sensing Framework for IoT

A short review for "An Attention Mechanism Inspired Selective Sensing Framework for Physical-Cyber Mapping in Internet of Things"

Edited by Jinbo Xiong

Ning H, Ye X, Sada A B, et al. An Attention Mechanism Inspired Selective Sensing Framework for Physical-Cyber Mapping in Internet of Things[J]. IEEE Internet of Things Journal, 2019, 6(6): 9531-9544.

With the increasing demand of ubiquitous computing, services, and intelligence in IoT, ubiquitous sensing as the primary premise of acquiring data has received widespread attention [1]-[3]. Benefiting from the growing advances of sensors in miniaturization, cost, and power, recent years have witnessed a large-scale deployment of sensors, which make it possible to better perceive the surroundings and provide automatic and accurate services for humans [4]-[6]. In this case, data overload and resource waste inevitably become two key problems in terms of data sensing. On one hand, the dynamic and uncertain environment boosts the explosive growth of data at a rapid pace while the sensing resources (e.g. sensors, communication, and storage) being relatively limited. The bottleneck between the explosive big data and limited sensing resources is no doubt challenging the ubiquitous sensing of IoT in the future. On the other hand, the full processing of raw sensor data not only requires large amounts of computing resources but also may lead to uncertain and unnecessary resource waste since some raw data may be worthless for later processing.

Selective sensing is thus proposed to tackle problems and it attracts wide attention with its unique advantages. As its name implies, it refers to selectively processing raw sensor data based on the real-time situation instead of attempting to fully process the massive data in parallel. However, most of current solutions are either passive relying on users' requests or narrow-application with great limitations. A more general and effective solution for large-scale IoT application is necessary.

Generally, human or animals can fleetly select the information of interest from a complicated and noisy environment. In other words, only the incoming sensory information that we want or need are allowed to reach the short-term memory and higher levels of information processing in the brain [7]-[9]. This act of information selection is often referred to as the attention mechanism. Inspired by the remarkable data processing capability of creatures, a new attention mechanism inspired selective sensing (AMiSS) framework was proposed in this paper, aiming to provide a general solution for the sensing bottleneck and resource waste in large-scale IoT application.

Drawing on the four sub-components of attention mechanism, i.e. focused attention, sustained attention, selective attention, and divided & alternating attention [10]-[13], the AMiSS was designed into five main components: attention building, attention maintaining, attention diversion, attention withdrawal, and associated memories pool and rule library (AMPRL).

The objective of the attention building is to catch and concentrate the attention of the system on a specific task or event, which contains selective filtration and multi-mode attention fusion. Particularly, two strategies of top-down control and bottom-up effect of biological attention mechanism were introduced in AMiSS, corresponding to task-guided attention and data-driven attention. Moreover, the biased competition based on the saliency of stimuli was referred to determine the importance of different sensed data in relevance to the target and distinguish the data of interest from a large number of irrelevant data.

The attention maintaining was designed to hold and keep the focused attention over a prolonged period of time. In analogy to the biological sustained attention, this component provides two sub-functions, concentration and vigilance. The former aims to continuously track the task-based attention window (AW) to guarantee the successful performance of the attention tasks,

while the latter is set for detecting new saliency in the scope of focused attention. Besides, resources management & optimization was also designed as a key module for attention maintaining.

The attention diversion occurs when a new task is issued or a higher saliency data is received. Both cases correspond to the expectation-guided attention diversion and event-driven attention diversion, respectively. Supported by the analysis of stimulating factors (e.g. the importance, urgency, interestingness, and relevance of the task), the decision-making module determines whether shift or divide the attention to other tasks.

The purpose of attention withdrawal is to release the attention resources when no information requires to be processed (task-driven) or any abnormity happens (exception-driven). The AMPRL plays a similar role to the memory and incremental learning to support the decision and prediction of the system.

In summary, the proposed AMiSS framework is highly innovative, which also well adapts to the development needs of IoT. Although it is just a narrow demonstration with many limitations, the proof-of-concept simulation still shows that the feasibility and effectiveness of the AMiSS in practice, as well as its promising benefits in tackling the sensing bottleneck and resource waste in a large-scale IoT application.

References:

[1] H. Ning, X. Ye, M. A. Bouras, D. Wei, and M. Daneshmand, General cyberspace: Cyberspace and cyber-enabled spaces, IEEE Internet of Things Journal, 2018.

[2] L. Atzori, A. Iera, and G. Morabito, The internet of things: A survey, Computer networks, vol. 54, no. 15, pp. 2787 2805, 2010.

[3] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, Context aware computing for the internet of things: A survey, IEEE communications surveys & tutorials, vol. 16, no. 1, pp. 414 454, 2014.

[4] J. Ma, “Smart u-things–challenging real world complexity,” in IPSJ symposium series, vol. 19, 2005, pp. 146–150.

[5] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelffl´e, “Vision and challenges for realising the internet of things,” Cluster of European Research Projects on the Internet of Things, European Commision, p.229, 2010.

[6] J. Gao, L. Lei, and S. Yu, Big data sensing and service: a tutorial, in Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on. IEEE, 2015, pp. 79 88.

[7] G. Billock, C. Koch, and D. Psaltis, Selective attention as an optimal computational strategy, in Neurobiology of Attention. Elsevier, 2005, pp. 18 23.

[8] L. Itti, Models of bottom-up attention and saliency, in Neurobiology of attention. Elsevier, 2005, pp. 576 582.

[9] S. Grossberg, Linking attention to learning, expectation, competition, and consciousness, in Neurobiology of attention. Elsevier, 2005, pp. 652 662.

[10] J. R. Sternberg and K. Sternberg, “Cognitive psychology,” Science, p.609, 2011.

[11] G. Daniel and I. Cristina, Attention. American Cancer Society, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470018860.s00481.

[12] T. P. P. Center, Types of attention, http://thepeakperformancecenter.com/educationallearning/learning/process/obtaining/types-of-attention/, accessed May 17, 2018.

[13] MindMed, Understanding the 4 types of attention, http://www.adhdapp.com/2013/02/06/understanding-the-4-types-of-attention/, accessed May 17, 2018.

Jinbo Xiong, Ph.D, is an Associate Professor in the Fujian Provincial Key Laboratory of Network Security and Cryptology and the College of Mathematics and Informatics at Fujian Normal University. He received his M.S. degree in Communication and Information Systems from Chongqing University of Posts and Telecommuni-cations, China, in 2006, and Ph.D. degree in Computer System Architecture from Xidian University, China, in 2013. His research interests include cloud data security, mobile data management, privacy protection, and Internet of Things. He has published papers in prestigious journals such as IEEE Internet of Things Journal, IEEE Transactions on Cloud Computing, and in major International conferences such as IEEE ICCCN, IEEE TrustCom, IEEE HPCC and IEEE ICPADS. He is a member of IEEE.

Symmetric ICP for Point Cloud Alignment

A short review for "A symmetric objective functions for ICP"

Edited by Dr. Carsten Griwodz

Szymon Rusinkiewicz, "A symmetric objective functions for ICP," ACM Transactions on Graphics, 38(4), t pages, DOI: 10.1145/3306346.3323037

Volumetric data receives increasingly the attention of the multimedia community. It may actually describe the content of a volume or the surfaces of objects. The first category can be described by a sequence of images that describe a space in terms of layered images [1] or potentially by point clouds [2]. The description of a scene in terms of surfaces may not seem to fit the term "volumetric" data equally well, but it is actually the more frequent kind of data. It can be represented by collections of images, meshes, or point clouds.

In the majority of cases, high-quality volumetric data requires several independent recordings that must be merged to improve quality. While images are arranged either during recording or through the use of various image registration techniques, meshes and point clouds require different methods. One of the foremost techniques for merging several point clouds relies on the registration technique Iterative Closed Points (ICP).

ICP algorithms have existed for several decades and become prominent since Besl and McKay [3] proposed an iterative approach that could align point sets both in terms of 3D translation and rotation. However, with the advent of volumetric data in multimedia research, they become rather relevant also for the multimedia community. Their most important function is the matching of various sets of volumetric data whose position and orientation in space is unknown, and which can potentially also have arbitrary differences in scale.

ICP algorithms try to solve the problem of aligning two point clouds which partially or completely present the same scene, but which are in different coordinate systems. The algorithms rotate, translate, and scale one of the clouds until it is sufficiently close to the other one by some metric. The cloud which is transformed is called source, the fixed cloud is the target.

The original ICP algorithm is a point-to-point algorithm. The objective function of point-to-point algorithms finds a transformation consisting of rotation and translation that minimizes the sum of the squared distance between every point in one transformed source point cloud and its closest point in the target point cloud.

Chen and Medioni [4] had proposed an objective function for matching surfaces that was adopted in the point-to-plane algorithms for ICP. In this case, the surface normals in all points of a point cloud are computed (as the smallest Eigenvector of the covariance matrix of the points surrounding point each point), and the distances between every point in the transformed source point cloud and the closest tangent planes in the target point cloud are minimized.

For point clouds that fill volumes instead of filling surfaces, point-to-plane ICP is not suitable because the creation of normals for points inside a volume would invent structures that do not exist. For point clouds that describe surfaces, however, point-to-plane algorithms exhibit faster convergence than point-to-point algorithms [5]. Also, if points have imperfect positions, point-to-plane algorithms do generally have superior matching results.

The new paper "A Symmetric Objective Function for ICP" by Szymon Rusinkiewicz proposes a variation of the existing point-to-plane algorithms. The main difference from existing work lies in the objective function that is used to minimize the distance between two point clouds. Symmetric ICP does not contribute any work on estimating scale differences between point clouds.

The title hints at its properties: it is symmetric because it computes the distance between a pair of points p and q in source and destination, respectively multiplied by the sum of the surface normal in both points n_p and n_q, and it the results is 0, then p and q are located on a common plane. Searching for solutions in 4D space gives a lot of freedom for transforming the source, but sensible translation and rotation opportunities that minimize the distance are directly available from a translation vector and the rotation. An important point is that both points are rotated by the same amount but in opposite directions. By expressing

the rotations as a Rodrigues rotation, the least-squares solution is solved for the rotation vector and a rotated translation. The paper does then show that a solution can deal with translations along the Z-axis better than other solutions because the multiplication of the vector *p-q* with the sum of normals approaches zero even in case of a translation in the Z direction. Furthermore, Rusinkiewicz points out that the exact solution can be found for the linearized rotations because the solution does not rely on solving any Eigenvector problem. On top of this, Rusinkiewicz can demonstrate a convergence rate that increases as the error shrinks.

This theoretical paper improves the state-of-the-art in ICP, although the author leaves separate points open for future analysis to understand and explain the good performance of Symmetric ICP better.

He does, however, go one step further and offers an open-source implementation [1] of Symmetric ICP that is integrated in the C++ library Trimesh 2. Practical testing shows that Symmetric ICP improves on other existing open-source implementations of ICP both in terms of speed and accuracy.

In the paper, it is discussed that partially overlapping point clouds need would benefit from one of the existing advanced robust estimation techniques rather than the outlier rejection approach that is integrated into the reference implementation for the paper.

While this is certainly a that should be done in the future, the approach is already working very well for finding the best transformation between point clouds that cover the same area but with strongly diverging densities.

We found that Symmetric ICP is a tool that provides an excellent tool for merging several LiDAR scans of indoor environments that consist only of points without additional attributes, where the base coordinate system can be arbitrarily rotated and translated between scans, while the scale is fixed. Because of the nature of rotating LiDAR scanners, point densities vary widely. Symmetric ICP tolerates this very well. Also, the fact that the LiDAR scanner had a depth estimation accuracy limited to 3 centimeters (implying considerable deviations in surface normals) did not reduce the success of Symmetric ICP to find a good solution. In fact, the measured depth error distribution of individual scans of a flat wall was identical to the depth error distribution of scans that were merged using Symmetric ICP. Also, in using Symmetric ICP for a SLAM-based technique that made use of point-clouds fragments (snapshots) captured by a slowly moving LiDAR in the same kind of environment could be used to create the robust initial alignment for merging point clouds of regions of indoor space with very limited overlap.

Symmetric ICP is thus a very valuable tool in the toolbox of multimedia researchers who intend to work with point cloud data that lacks a attributes aside from 3D point positions. It converges very fast from a decent initial position and tolerates inaccurate point positions very well. Also Symmetric ICP's ability to work with point clouds that have strongly variable point densities throughout a scan and match them well with other points clouds of rather different densities is a strong point.

[1] https://gfx.cs.princeton.edu/pubs/Rusinkiewicz_2019_ASO/index.php

References:

[1] M. T. Orchard, A. Nosratinia, and R. Rajagopalan, "On interframe coding models for volumetric medical data," in *Proceedings., International Conference on Image Processing*, vol. 2, pp. 17–20.

[2] M. Schwarz and H.-P. Seidel, "Fast parallel surface and solid voxelization on GPUs," in *ACM SIGGRAPH Asia 2010 papers on - SIGGRAPH ASIA '10*, 2010, p. 1.

[3] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[4] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image Vis. Comput.*, vol. 10, no. 3, pp. 145–155, Apr. 1992.

[5] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat, "Comparing ICP variants on real-world data sets," *Auton. Robots*, vol. 34, no. 3, pp. 133–148, Apr. 2013.

Carsten Griwodz is professor at the University of Oslo. His research interest is the performance of multimedia systems. He is concerned with streaming media, which includes all kinds of

media that are transported over the Internet with a temporal demands, including stored and live video as well as games and immersive systems. To achieve this, he wants to advance operating system and protocol support, parallel processing and the understanding of the human experience. He was area chair of ACM MM 2019 and 2014, and general chair of ACM MMSys and NOSSDAV (2013), co-chair of ACM/IEEE NetGames (2011), NOSSDAV (2008), SPIE/ACM MMCN (2007) and SPIE MMCN (2006), TPC chair ACM MMSys (2012), and systems track chair ACM MM (2008). More information can be found at http://mlab.no

Paper Nomination Policy

Following the direction of MMTC, the Communications – Review platform aims at providing research exchange, which includes examining systems, applications, services and techniques where multiple media are used to deliver results. Multimedia includes, but is not restricted to, voice, video, image, music, data and executable code. The scope covers not only the underlying networking systems, but also visual, gesture, signal and other aspects of communication. Any HIGH QUALITY paper published in Communications Society journals/magazine, MMTC sponsored conferences, IEEE proceedings, or other distinguished journals/conferences within the last two years is eligible for nomination.

Nomination Procedure

Paper nominations have to be emailed to Review Board Directors: Qing Yang (qing.yang@unt.edu), Roger Zimmermann (rogerz@comp.nus.edu.sg), Wei Wang (wwang@mail.sdsu.edu), and Zhou Su (zhousu@ieee.org). The nomination should include the complete reference of the paper, author information, a brief supporting statement (maximum one page) highlighting the contribution, the nominator information, and an electronic copy of the paper, when possible.

Review Process

Members of the IEEE MMTC Review Board will review each nominated paper. In order to avoid potential conflict of interest, guest editors external to the Board will review nominated papers co-authored by a Review Board member. The reviewers' names will be kept confidential. If two reviewers agree that the paper is of Review quality, a board editor will be assigned to complete the review (partially based on the nomination supporting document) for publication. The review result will be final (no multiple nomination of the same paper). Nominators external to the board will be acknowledged in the review.

Best Paper Award

Accepted papers in the Communications – Review are eligible for the Best Paper Award competition if they meet the election criteria (set by the MMTC Award Board). For more details, please refer to http://mmc.committees.comsoc.org/.

MMTC Communications – Review Editorial Board

Multimedia Communications Technical Committee Officers

MMTC examines systems, applications , services and techniques in which two or more media are used in the same session. These media include, but are not restricted to, voice, video, image, music, data, and executable code. The scope of the committee includes conversational, presentational, and transactional applications and the underlying networking systems to support them.