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

MMTC Communications – Review

Vol. 13, No. 3, June 2022

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the June 2022 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises five reviews that cover multiple facets of multimedia communication research including Visual and audio information, Video Quality, the visual quality of videos, and multi-view video compression. These reviews are briefly introduced below.

The first paper, published in IEEE Transactions on Multimedia, Vol. 23, 2021. and edited by Dr. Takuya Fujihashi, focuses on Visual and audio information.

The second paper is published in IEEE Trans. Multimedia, vol. 24, pp. 1-16, 2022 and edited by Dr. Mengbai Xiao. It studies the worker selection issues to Accurately assessing the visual quality of videos.

The third paper, published in IEEE Transactions on Multimedia and edited by Dr. Edited by Dr. Jinbo Xiong, focuses on accurately assessing the visual quality of videos

The fourth paper, published in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3052183 and edited by Prof. Qichao Xu. This paper proposes an anonymous and privacy-preserving federated learning scheme.

The fifth paper, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 12, December 2021 Edited by Dr. Debashis Sen. The paper being reviewed here surveys several methods to evaluate visual scanpath prediction.

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

Zhisheng Yan
George Mason University, USA
Email: zyan4@gmu.edu

Yao Liu
Rutgers University, USA
Email: yao.liu@rutgers.edu

Wenming Cao
Shenzhen University, China
Email: wmcao@szu.edu.cn

Phoenix Fang
California Polytechnic State University, USA
Email: dofang@calpoly.edu/span>

Perceptual Vibrotactile-Signal Compression Based-on Sparse Linear Prediction

A short review for "PVC-SLP: Perceptual Vibrotactile-Signal Compression Based-on Sparse Linear Prediction"

Edited by Takuya Fujihashi

R. Hassen. B. Gülecyüz, E. Steinbach, "PVC-SLP: Perceptual Vibrotactile-Signal Compression Based-on Sparse Linear Prediction," IEEE Transactions on Multimedia, Vol. 23, 2021.

Visual and audio information are predominant in multimedia systems. Many technical solutions have been designed for the acquisition, storage, transmission, and display of these modalities. In recent years, the sense of touch (also called haptic information) is considered a new modality for enabling remote physical interaction with convincing touch experiences, i.e., Tactile Internet [1]. However, the acquisition, compression, transmission, and display of haptic information have not yet reached the same level as the solutions for the visual and audio information. In this paper, the authors aim to compress the haptic information. Note that haptic information consists of two submodalities, i.e., kinesthetic and tactile information, and the discussion of this paper covers tactile signal compression.

The existing tactile compression techniques can be categorized into two groups. The first group is based on transform coding. The typical solution is Haptic Codec [2]. The tactile signal is represented as the vibrotactile signal obtained from the three-axis acceleration sensor. The vibrotactile signal is transformed into a different domain using orthogonal transforms such as discrete cosine transform (DCT) or discrete wavelet transform (DWT). The domain representations are then quantized and entropy coded for compression. Here, Haptic Codec adopts a psychohaptic model [3] for non-uniform quantization to reduce perceivable distortion.

The second group is based on analysis-by-synthesis techniques. Specifically, the vibrotactile signal is synthesized using only the filter parameters during the analysis phase. The most common analysis-by-synthesis technique is linear predictive coding (LPC), which was mainly used in speech compression, and the existing studies [4] adopted LPC to minimize the mean-squared-error between the original and reconstructed signals.
In this paper, the authors present a hybrid approach of the two groups, PVC-SLP, for further efficient vibrotactile signal compression. The key contributions of this paper are two-fold: 1) PVC-SLP adopts analysis-by-synthesis and transform coding in sequence to fully utilize both advantages, and 2) perceptual quantization of the residual signal based on a cutaneous sensitivity model [5] to retain a lofty level of perceptional quality.

In PVC-SLP, they assume that vibrotactile signal can be modeled as a stationary random process using the autoregressive model. They adopt sparse linear prediction (SLP) to estimate AR model coefficients that minimize the L1-based cost function based on the model. In contrast to least squares and L2 solutions, they found L1-based solution is robust to noise as it tends to reject outliers. Since the model coefficients are needed to share with the receiver, PVC-SLP compresses the model parameters in addition to the vibrotactile signal. For the compression, it converts the model coefficients into reflection coefficients before quantization using Levinson-Durbin recursion because the reflection coefficients are robust to quantization error compared with the direct quantization of the model coefficients. The quantized reflection coefficients compute the residual signal between the original and predicted vibrotactile signal.

The residual signal is then decomposed using 1D-DCT. Although the DCT coefficients bring high compression gains, PVC-SLP aims to minimize the perceptual distortion. For this purpose, they adapt the quantization scheme to the cutaneous sensitivity model. The cutaneous sensitivity model is called acceleration sensitivity function (ASF), and the range is normalized into 0 through 1. A small value of ASF represents that the corresponding DCT coefficient will not impact the perceptual quality, and thus PVC-SLP discards the coefficient at the quantization. Finally, the quantized coefficients are entropy coded based on zero run length (ZRL) and Huffman coding.

The authors carried out experimental evaluations to clarify the coding efficiency and complexity using vibrotactile databases, which are the TUM database and LMT Haptic Material database. They adopt quality metrics of the vibrotactile signal, including signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR), and ST-SIM [6], to discuss the reconstruction quality.

The authors firstly discuss the baseline performance of PVC-SLP under the different haptic databases and quantization scales for the residual signal and reflection coefficients. For example, the average PSNR performance of TUM database is higher than that of LMT database. This is because of the wider diversity of vibrotactile signal content in LMT database. The authors also compared the compression performance against lossy and lossless codecs. The authors highlight the evaluation results as follows:

- The proposed quantization based on ASF maintains high perceived signal quality under the same compression ratio as the lossy codec.
- The lossless codec causes a large compression ratio. Lossy vibrotactile codecs bring a benefit, especially in media storage and transmission.

References:

[1] M. Simsek, A. Aijaz, M. Dohler, J. Sachs and G. Fettweis, "5G-Enabled Tactile Internet," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 3, pp. 460-473, March 2016.

[2] A. Noll, B. Gülecyüz, A. Hofmann and E. Steinbach, "A Rate-scalable Perceptual Wavelet-based Vibrotactile Codec," IEEE Haptics Symposium (HAPTICS), 2020, pp. 854-859.

[3] R. Chaudhari, C. Schuwerk, M. Danaei, and E. Steinbach, "Perceptual and bitrate-scalable coding of haptic surface texture signals," IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 3, pp. 462–473, 2014.

[4] R. Chaudhari, C. Schuwerk, M. Danaei, and E. Steinbach, "Perceptual and bitrate-scalable coding of haptic surface texture signals," IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 3, pp. 462–473, April 2015.

[5] S. J. Bolanowski Jr, G. A. Gescheider, R. T. Verrillo, and C. M. Checkosky, "Four channels mediate the mechanical aspects of touch," The Journal of the Acoustical Society of America, vol. 84, no. 5, pp. 1680–1694, 1988.

[6] R. Hassen and E. Steinbach, "Vibrotactile signal compression based on sparse linear prediction and human tactile sensitivity function," in IEEE World Haptics Conference (WHC), 2019.

Takuya Fujihashi received the B.E. degree in 2012 and the M.S. degree in 2013 from Shizuoka University, Japan. In 2016, he received Ph.D. degree from the Graduate School of Information Science and Technology, Osaka University, Japan. He is currently an assistant professor at the Graduate School of Information Science and Technology, Osaka University since April, 2019. He was an assistant professor at the Graduate School of Science and Engineering, Ehime University, Japan from Jan. 2017 to Mar. 2019. He was research fellow (PD) of Japan Society for the Promotion of Science in 2016. From 2014 to 2016, he was research fellow (DC1) of Japan Society for the Promotion of Science. From 2014 to 2015, he was an intern at Mitsubishi Electric Research Labs. (MERL) working with the Electronics and Communications group. He received best paper award in IEEE ICCE 2022 and selected one of the Best Paper candidates in IEEE ICME (International Conference on Multimedia and Expo) 2012. His research interests are in the area of video compression and communication.

A Spatial-Temporal Distortion Metric for Assessing 360-Degree Video Quality

A short review for "Quality Assessment for Omnidirectional Video: A Spatio-Temporal Distortion Modeling Application"

Edited by Mengbai Xiao

P. Gao, P. Zhang, and A. Smolic, "Quality Assessment for Omnidirectional Video: A spatial-Temporal Distortion Modeling Application," in IEEE Trans. Multimedia, vol. 24, pp. 1-16, 2022.

Accurately assessing the visual quality of videos is not easy. The video quality assessment (VQA) metrics are either objectively or subjectively measured. For the objective VQA metrics, the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are the most widely adopted ones. But they are not accurately aligned with the watching experience of the users. On the other hand, the researchers could recruit a group of people to watch and evaluate the video content, and the quality of a video is characterized with the mean opinion score (MOS). However, collecting the subjective VQA metrics consumes substantial time and resources, and the results might be inconsistent because people are easily affected by their mood and the environment.

When focusing on the emerging omnidirectional videos, assessing the quality of videos is more challenging. Pixels are projected and anti-projected between a virtual sphere for watching and a rectangular plane for coding. Thus, PSNRs and SSIMs that are only calculated on the 2D planes are not guaranteed to precisely summarize the visual distortion happening on the sphere when watching. Under this circumstance, the objective VQA metrics are extended to S-PSNR [2], CPP-PSNR [3], and WS-PSNR [4]. However, the prior studies incorporate the spatial distortion only, ignoring that the human visual system (HVS) is also affected at the temporal dimension.

In this paper, Gao et. al. [1] propose to develop a novel VQA metric that captures the temporal change of spatial distortion in omnidirectional videos, namely OV-PSNR. OV-PSNR is a full-reference (FR) method, so it is evaluated between an original video sequence and an impaired video sequence. OV-PSNR consists of four steps to derive the expected spatio-temporal distortion for a video: 1) spatio-temporal tube creation, 2) spatial distortion evaluation, 3) spatio-temporal tube distortion evaluation, and 4) pooling. The spatio-temporal tube is directly created on the 2D planes where the block division and motion estimation are less difficult. Specifically, the spatio-temporal tubes are created on a number of consecutive frames. A frame in the group is divided into blocks, and for each block, the temporally correlated blocks in its previous frames are recursively discovered via motion estimation. These temporally correlated blocks are combined to form a spatio-temporal tube. In the second step, the spatial distortion is evaluated between blocks of the original and the impaired video sequences. Various metrics representing the spatial distortion [2, 3, 4] could be incorporated. As long as a spatio-temporal tube has been created and spatial distortion of the blocks has been calculated, a distortion score could be summarized for the tube. The score of tube is derived from both the average distortion and the temporal distortion. The average distortion is calculated as a smoothed value by recursively summing the block-wise distortion, where the distortion gradient of consecutive frames determines two ways of computation. The temporal distortion represents the combination of the amplitude and frequency of the distortion gradient. In the last stage, the distortion score of a frame is calculated by pooling the distortion scores of tubes ending at this frame, and all frame-level distortion scores are summarized to form the OV-PSNR.

In the calculation pipeline of OV-PSNR, various metrics could be involved to calculate the block-wise distortion map. However, the integration of WS-PSNR, S-PSNR, and CPP-PSNR is not the same. If WS-PSNR is adopted, the weights of pixels are directly used in calculating the distortion of 2D blocks, but this requires that the orginal video sequence and the impaired video sequence share the same projection format and the same resolution. When using S-PSNR, the pixels on an intermediate sphere bridge the pixels in the original and the impaired blocks. The nearest neighboring search based on K-D tree is applied to solve the mismatching between pixels from different domains. CPP-PSNR is calculated also by adding an intermediate domain except that the CPP domain is used instead of the spherical domain. It worth noting that with the use of a unifying intermediate domain, S-PSNR and CPP-PSNR both supports to generate the distortion map for

blocks with different projection formats, e.g., equal-area projection (ERP), icosahedron projection (ISP), octahedron projection (OHP), etc.

The authors evaluate OV-PSNR on the VR-VQA48 dataset [5]. To justify the proposed metric, the experiments are carried out to measure if OV-PSNR is consistent to the subjective metric, different MOS (DMOS). The results show that OV-PSNR and its variants substantially outperform other objective distortion metrics, which means OV-PSNR more accurately characterize the user's subjective feeling on the omnidirectional videos. Among the OV-PSNR variants, the one using WS-PSNR achieves the best performance because it measures quality from the original signals, i.e., not intermediate domain introduced. In terms of the execution time, generally the OV-PSNR family consumes more time because of the high complexity of computing temporal distortion, especially the tube creation part. Another set of experiments on VQA-ODV dataset [6] also verify the conclusion that OV-PSNR and its variants are the most consistent objective metrics to the subjective metrics.

In summary, the authors of this paper design an objective metric, namely OV-PSNR, that assesses the omnidirectional video quality. Compared to the state-of-the-art metrics, OV-PSNR captures the distortion at the temporal dimension and is more consistent to the subjective metrics, e.g., DMOS, though additional computation resources have to be devoted. I expect this metric would widely replace the currently deployed ones in a wide range of scenarios, e.g., video coding and streaming, and substantially benefits the omnidirectional video techniques.

References:

[1] P. Gao et al., “Quality Assessment for Omnidirectional Video: A Spatio-Temporal Distortion Modeling Approach,” in IEEE Transactions on Multimedia, vol. 24, pp. 1-16, 2022.

[2] M. Yu et al., “A Framework to Evaluate Omnidirectional Video Coding Schemes,” in Proc. IEEE Int. Symp. Mixed Augmented Reality, pp. 31-36, 2015.

[3] V. Zakharchenko et al., “Quality Metric for Spherical Panoramic Video,” Opt. Photon. Inf. Process. X, vol. 9970, pp. 57-65 2016.

[4] Y. Sun et al., “Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video,” IEEE Signal Process. Lett., vol. 24, no. 9, pp. 1408-1412, 2017.

[5] VR-VQA48 database, with subjective test data for panoramic video quality and head tracking data. Available: https://github.com/Archer-Tatsu/head-tracking

[6] VQA-ODV database, a large-scale dataset of omnidirectional video for visual quality assessment. Available: https://github.com/Archer-Tatsu/VQA-ODV

Mengbai Xiao, Ph.D., is a Professor in the School of Computer Science and Technology at Shandong University, China. He received the Ph.D. degree in Computer Science from George Mason University in 2018, and the M.S. degree in Software Engineering from University of Science and Technology of China in 2011. He was a postdoctoral researcher at the HPCS Lab, the Ohio State University. His research interests include multimedia systems, parallel and distributed systems. He has published papers in prestigious conferences such as ACM Multimedia, ACM ICS, IEEE ICDE, IEEE ICDCS, IEEE INFOCOM.

3D Point Clouds Feature Fusion-Based Vehicle Cooperative Perception

A short review for "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds"

Edited by Jinbo Xiong

Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, New York, NY, USA, pp. *88-100, Nov. 2019,* DOI: 10.1145/3318216.3363300.

Autonomous vehicles rely heavily on sensors to improve the perception of their surroundings. When the sensors fails, or due to their own performance limitations, such as the object is blocked, truncated, too far away, etc. Object perception and detection will be very bad. Connected autonomous vehicles (CAVs) offer a promising solution that can fuse the perception data of multiple vehicles to achieve a larger range and more accurate object detection [1]. Generally, object data fusion can be divided into signal-level, feature-level and detection-level [2]. Although the detection-level fusion method has a high speed and only needs to share the detection results of each vehicle, it cannot detect the object that is not detected by a single vehicle [3]. In order to enhance semantic information exchange of perception data, Chen et al. proposed Cooper, which directly fused raw LiDAR point cloud data from multiple vehicles. However, this low-level method requires large data transmission, which poses challenges to channel bandwidth and processing delay [4]. In addition, the sharing of raw data also leads to the disclosure of private information [5].

In order to solve the above problems while fully developing and utilizing the semantic information of feature map, in this paper, the authors propose a feature-level data fusion method to realize vehicle cooperative perception. The proposed F-Cooper framework uses VoxelNet network [6] to extract the voxel features of raw point cloud, and realizes voxel feature fusion and spatial feature fusion to improve the accuracy of object detection. Similar to raw point cloud fusion, the feature fusion can be decoupled into two aspects of position alignment and feature selection. Compared to Cooper, F-Cooper requires only one hundredth of the size of original data and can be deployed and executed in on-board and on-edge. Additionally, the sharing of feature map effectively reduces privacy concerns. Specific contributions devoted by the authors can be summarized into the following three aspects.

Firstly, in order to prove the correctness and rationality of feature fusion, the authors analyze some basic conditions and assumptions. Compared with the original data, the feature map filters out the data irrelevant to the object detection and only preserves the valid semantic and position information. Benefit from the modern autonomous vehicle object detection technology, convolutional neural network (CNN) is mostly used to process the original data, and the extracted feature mas have similar data types and formats. position alignment and feature fusion can be achieved by sharing GPS and IMU data between vehicles.

Second, the authors design a voxel feature fusion (VFF) method that employs VoxelNet to group and encode point cloud into multi-dimensional voxel features. The relationship between the voxel points includes four positional relationships, namely, falls within voxel, falls on a side of voxel, falls on an edge of voxel, and falls on a corner point of voxel. The voxel feature vectors at the same position were fused using the maxout method. The served voxel features can better express the object semantics.

Thirdly, the authors design a spatial feature fusion (SFF) method. Using the GPS data of the interactive vehicles, the relative position and driving angle of the sending vehicle and the receiving vehicle are aligned and calibrated. All vehicles share a global 3D coordinate system. According to the sequence of feature channels, the non-overlapping features are retained, and the overlapping features are fused by maxout operation. Additionally, since spatial features are sparse, feature compression can be used to further improve the efficiency of channel transmission and feature fusion.

Thereafter, with Tom & Jerry (T&J) dataset consisting a lot of 16-beam point cloud samples, the authors conduct comprehensive experiments in the three scenarios of road intersections, multi-lane roads and campus parking lots. Experimental results show that F-Cooper can

almost achieve the same detection accuracy compared with Cooper. Also, F-Cooper effectively reduce the size of transmitted data. Each LiDAR frame data contains about one hundred thousand points or about 4 MB, while the features extracted by CNN only about 200 KB. These features can be transmitted in tens of milliseconds, making real-time vehicle data fusion and cooperative perception possible.

In summary, the proposed F-Cooper framework effectively improve the communication burden and privacy concerns of signal-level fusion schemes. The authors make a breakthrough in the approach of point cloud feature fusion, which provides a promising solution for low-delay vehicle cooperation perception.

References:

[1]Q. Yang, S. Fu, H. Wang and H. Fang, "Machine-Learning-Enabled Cooperative Perception for Connected Autonomous Vehicles: Challenges and Opportunities," in IEEE Network, vol. 35, no. 3, pp. 96-101, Jun. 2021, DOI: 10.1109/MNET.011.2000560.

[2]J. Shi, W. Wang, X. Wang, H. Sun, X. Lan, J. Xin and N. Zheng, “Leveraging Spatio-Temporal Evidence and Independent Vision Channel to Improve Multi-Sensor Fusion for Vehicle Environmental Perception,” in Proceedings of the IEEE Intelligent Vehicles Symposium, Changshu, China, pp. 591-596, Jun. 2018, DOI: 10.1109/IVS.2018.8500665.

[3]Q. Chen, S. Tang, J. Hochstetler, J. Guo, Y. Li, J. Xiong, Q. Yang and S. Fu, “Low-Latency High-Level Data Sharing for Connected and Autonomous Vehicular Networks,” in Proceedings of the IEEE International Conference on Industrial Internet, Orlando, FL, USA, pp. 287-296, Nov. 2019, DOI: 10.1109/ICII.2019.00055.

[4]Q. Chen, S. Tang, Q. Yang and S. Fu, “Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds,” in Proceedings of the IEEE 39th International Conference on Distributed Computing Systems, Dallas, TX, USA, pp. 514-524, Jul. 2019, DOI: 10.1109/ICDCS.2019.00058.

[5]J. Xiong, R. Bi, M. Zhao, J. Guo and Q. Yang, “Edge-Assisted Privacy-Preserving Raw Data Sharing Framework for Connected Autonomous Vehicles,” in IEEE Wireless Communications, vol. 27, no. 3, pp. 24-30, Jun. 2020, DOI: 10.1109/MWC.001.1900463.

[6]Y. Zhou and O. Tuzel, “Voxelnet: End-to-end learning for point cloud based 3d object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4490-4499, Jun. 2018, DOI: 10.1109/CVPR.2018.00472.

Jinbo Xiong, is a Professor and Ph.D. supervisor with the Fujian Provincial Key Laboratory of Network Security and Cryptology and the College of Computer and Cyber Security at Fujian Normal University. He received the M.S. Degree in Communication and Information Systems from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2006, and the Ph.D. degree in Computer System Architecture from Xidian University, China, in 2013. He was a visiting scholar with the Department of Computer Science and Engineering at the University of North Texas, Denton, USA, from 2019-2020. His research interests include connected and autonomous vehicle, secure deep learning, cloud data security, mobile media management, privacy protection, and Internet of Things. He has published more than 100 papers in prestigious journals such as IEEE WCM, IEEE TII, IEEE TCC, IEEE IoT J, IEEE TNSE, FGCS and in major International conferences such as IEEE ICCCN, IEEE TrustCom, IEEE HPCC and IEEE ICPADS. He has applied 15 patents and two monographs in these fields. He is a member of IEEE and senior member of CCF.

Anonymous and Privacy-Preserving Federated Learning With Industrial Big Data

A short review for "Anonymous and Privacy-Preserving Federated Learning With Industrial Big Data"

Edited by Qichao Xu

B. Zhao, K. Fan, K. Yang, Z. Wang, H. Li and Y. Yang, "Anonymous and Privacy-Preserving Federated Learning With Industrial Big Data," in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3052183.

With the rapid development of the Internet of Things and big data, the Industrial Internet has become the development trend of modern industry. Smart factories equipped with advanced sensors can generate and collect large amounts of real-time industrial big data during the production process.

Analysis of industrial big data through AI (AI, artificial intelligence) technology can improve system performance, reduce downtime, make easier to maintain and more[1]. In comparison with general big data, industrial big data has the multisource and heterogeneous nature, which makes traditional signal processing techniques no longer meet the requirements of data analysis. Moreover, industrial big data comprises some privacy-sensitive data, such as personal employee information and sensitive customer information, which may be leaked to manufacturers or service personnel of industrial equipment during the data analysis process[2]. An emerging deep learning method called Federated learning(FL) is suitable for multisource big data, which enables distributed parties to collaboratively train a model without data sharing. FL is more secure as only the model parameters are shared and not the raw data.

However, researches in recent years indicate that FL still exists various privacy issues[3]. Firstly, in the process of communication, the privacy of participants' identity information may be compromised. Besides, some malicious entities can infer sensitive information about participants' private training data through shared parameters[4], which results in the leakage of participants' privacy. Therefore, data protection mechanisms need to be introduced to address data security and privacy issues.

To solve to above problems, an anonymous and privacy-preserving federated learning scheme is proposed for mining the industrial big data. To be more specific, the privacy leakage is first reduced by sharing fewer parameters between participants and the server in FL. And the effect of the proportion of shared parameters on accuracy is experimentally explored. Then, the model parameters are perturbed utilizing differential privacy with the Gaussian mechanism to achieve strict privacy preservation. In addition, a fully trusted proxy server is introduced as an intermediate layer between participants and the server to achieve anonymity of participants while reducing the communication cost on the server during the learning process.

Therefore, in this article, the author's main contribution is to propose an anonymous and privacy-preserving federated learning scheme. In order to reduce the privacy leakage during FL, only partial participants and partial parameters are selected to update. In order to prevent the server from obtaining participants identities and reduce the communication cost on the FL server, a trustful proxy server is introduced as the middle layer in the scheme. And the Gaussian noise is added to the updated model parameters for strict privacy preservation.

The privacy-preserving FL system consists of multiple participants, a FL server, and a proxy server. In each iteration, the FL server first selects a subset of participants for training and distributes the global model to them. Participants download the global model and replace the corresponding parameters in the local model. Participants use the mini-batch gradient descent (MBGD) algorithm to minimize the loss function as a way to update the local model. The difference between the optimized local model and the global model is referred as participants' gradients. To prevent model parameters from leaking privacy, the Gaussian noise is added to the gradients. To reduce the potential privacy leakage, only partial updated gradients are chosen to be unchanged and the remaining gradients become zero. Then, all the gradients are sent to the proxy server. After the FL server receives all uploaded local model gradients from the proxy server, it aggregates them to get the updated global model.

To prevent the curious server from connecting the updates of each participant and reduce the communication burden on the FL server, a parameter server that ignores the identities of the uploaders is designed. And the trustful proxy server stores the addresses of the FL server and all participants. Both the FL server and the participants have their public-private key pairs and distribute their public keys in advance with

the help of a certificate authority (CA). The FL server randomly selects partial participants and encrypts the global model separately with their public keys. The encrypted parameters are uploaded to the proxy server and then forwards to the participants.

Since the anonymous communication, the server has no way of knowing who is involved in the training. Only the participants selected by the FL server can decrypt the parameters with their private keys. The selected participants optimize their local models and calculate the gradients. The Gaussian mechanism is leveraged on gradients to achieve DP for strict privacy preservation. Distortion governed by the noise variance is taken into account, once it exceeds a certain limit, the gradient information will be corrupted by the added noise, and accuracy of the global model will be affected. Thus a self-stop mechanism is set up by tracking the broken probability of DP and stop the whole training once it reaches a certain threshold. The moments accountant[5] is used to calculate the failure probability of DP, and it increases when the number of queries to gradients increases.

Extensive simulations have evaluated the performance of the proposed scheme. The simulation results show that the proposed scheme can achieve almost the same accuracy as unimproved FL when the shared parameters decrease by only one order of magnitude, while providing strict privacy preservation by sharing fewer parameters and adding Gaussian noise to shared parameters. Besides, the scheme had better robustness than traditional FL and DP-FL schemes, and it did not add too much computation cost.

In summary, this paper designed an anonymous and privacy-preserving FL system and applied an improved FL algorithm to the mining of industrial big data. First, privacy leakage is reduced by sharing fewer parameters, only partial participants and parameters were updated and shared in each iteration. Then, a proxy server was introduced to reduce the communication burden of the FL server and to prevent the server from obtaining the identities of participants. Moreover, DP with Gaussian mechanism was leveraged on updated gradients to provide strict privacy preservation. And a self-stop mechanism was established to prevent excessive data interference that may degrade model accuracy.

References:

J. Yan, Y. Meng, L. Lu, and L. Li, “Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance,” IEEE Access, vol. 5, pp. 23 484–23 4912017.

C. Yin, J. Xi, R. Sun, and J. Wang, “Location privacy protection based on differential privacy strategy for big data in industrial Internet of Things,” IEEE Trans. Ind. Informat., vol. 14, no. 8, pp. 3628–3636, Nov. 2017.

T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” IEEE Process. Mag., vol. 37, no. 3, pp. 50–60, May 2020.

M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., 2015, pp. 1322–1333.

M. Abadi et al., “Deep learning with differential privacy,” in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2016, pp. 308–318.

Qichao Xu, Ph.D, is an Assistant Professor in the School of Mechatronic Engineering and Automation, Shanghai University. He received the Ph.D. degree from Shanghai University, Shanghai, China, in 2019.

His research interests include Internet of Things, autonomous driving vehicles, and trust management. He has published more than 50 papers in prestigious journals such as IEEE TIFS, IEEE TMM, IEEE TITIS, IEEE TII, IEEE TVT, IEEE TBD and IEEE IoTs, in prestigious conferences such as IEEE ICC, IEEE INFOCOM.

Evaluation of Visual Scanpath Prediction Consistent with Human Subjective Assessment

A short review of "Evaluation of Saccadic Scanpath Prediction: Subjective Assessment Database and Recurrent Neural Network Based Metric"

Edited by Debashis Sen

Chen Xia , Junwei Han and Dingwen Zhang, " Evaluation of Saccadic Scanpath Prediction: Subjective Assessment Database and Recurrent Neural Network Based Metric," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 12, December 2021.

A visual scanpath refers to the human eye movements that scan the visual field to sample and receive visual information [1]. The main components of a visual scanpath are saccades and fixations, which repeat one after another playing central roles in overt human visual attention. While saccades are ballistic changes in eye position resulting in movement from one location to another, fixations after saccades bring the area of interest onto the high acuity fovea region of the retina for information extraction [2]. Prediction of visual scanpaths that represent the dynamic relationship of information at fixation locations are increasingly being considered as essential to model visual attention, particularly for image understanding [3]. Therefore, visual scanpath prediction for images is an important task to consider and it has recently received great research interest [1].

A visual scanpath on an image is often predicted as a sequence of image locations, which not only provides the important regions in the image but also their order. Several approaches for predicting visual scanpaths have been proposed in the last decade. While many of these approaches are built upon classical image saliency prediction models like [4, 5], the most recent ones such as [1, 3] are based on the potent use of deep learning. Most of these prediction models are often evaluated by computing some kind of similarity between the predicted visual scanpaths and all the corresponding human scanpaths considering a few publicly available datasets like [6, 7, 8].

The paper being reviewed here surveys several methods to evaluate visual scanpath prediction. It is found in the paper that the evaluation measures used in different works widely vary from each other due to lack of a unified standard. Some of the measures are found to arrive at conflicting conclusions as well. It is also observed that the conclusions drawn from some of the measures may not be consistent with human subjective evaluation.

In light of the above issues, the paper first studies the effectiveness of the different evaluation measures used in literature. A database is created containing 5000 pairwise comparisons of scanpaths using the OSIE eye-tracking dataset [7]. The human assessment of the similarity in the 5000 pairs of scanpaths is also collected. The subjective assessment dataset thus formed is used against the evaluation measures for the 5000 scanpath pairs to judge the effectiveness of the measures in terms of consistency with human subjective perception. It is concluded that there is a scope for improving consistency by considering data-driven evaluation and comparison measures for visual scanpath prediction.

In line with the suggestion to explore data-driven evaluation strategies, another major contribution of the paper being reviewed here is a data-based measure to compare two predicted scanpaths with reference to a ground truth human scanpath. The approach takes the two predicted scanpaths and the related reference scanpath as inputs, and outputs a binary decision indicating that one of the predicted scanpaths is more similar to the reference than the other.

First, a deep autoencoder network is used to learn a semantic encoding of a scanpath. Then, the semantic codes of the two predicted scanpaths and the reference scanpath are used to compute the dissimilarity between each pair of the predicted and reference scanpaths. The two quantified dissimilarities are then mapped to the binary decision declaring one of the two predicted scanpath to be closer to the reference scanpath. The mapping is performed using an LSTM network, which is trained on the human assessment data of the scanpath pairs formed to study the evaluation measures in literature.

The existing evaluation measures and the new data-driven evaluation approach are used in the paper to evaluate the state-of-the-art visual scanpath prediction models considering the Toronto [6], OSIE [7] and MIT1003 [8] data sets. Prediction models of three

categories, namely, saliency-based models, explicit dynamic saccade models, and baseline models (chance, center, and inter-observer models), are considered. It was concluded that the data-driven evaluation approach was less sensitive to biases such as image-center emphasis. It was also found that room remains to develop improved prediction models, as the baseline inter-observer model was quite ahead in performance to the models of the existing algorithms.

References:

[7] W. Sun, Z. Chen and F. Wu, "Visual Scanpath Prediction Using IOR-ROI Recurrent Mixture Density Network," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2101-2118, June 2021.

[8] O. Le Meur, A. Coutrot, Z. Liu, P. Rämä, A. Le Roch and A. Helo, "Visual Attention Saccadic Models Learn to Emulate Gaze Patterns From Childhood to Adulthood," in *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4777-4789, Oct. 2017.

[9] M. Assens Reina, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, "Saltinet: Scan-path prediction on 360 degree images using saliency volumes," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2331–2338, 2017.

[10] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in Proceedings of International Conference on Neural Information Processing Systems, pp. 545–552, 2006.

[12] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in Proceedings of International Conference on Neural Information Processing Systems, pp. 545–552, 2005.

[13] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," Journal of Vision, vol. 14, no. 1, pp. 1–20, Jan. 2014.

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2106–2113, 2009.

Debashis Sen is an Assistant Professor in the Department of Electronics and Electrical Communication Engineering and a faculty member in the Centre of Excellence in Advanced Manufacturing Technology of Indian Institute of Technology - Kharagpur. He received his Ph.D. in Image Processing from Jadavpur University, Kolkata, India and his M.A.Sc. in Electrical Engineering from Concordia University, Montreal, Canada. He was a postdoctoral researcher at the Multimedia Analysis and Synthesis Laboratory, National University of Singapore and at the Center for Soft Computing Research, Indian Statistical Institute. He currently heads the Vision, Image and Perception research group and the ArtEye Lab in his department, which are funded by multiple agencies of Government of India and prominent industries in India. His current research interests are in Vision, Image and Video Processing, Uncertainty Handling, Eye Movement Analysis, Machine Vision and Deep Learning. He has authored/co-authored more than 50 research articles in high impact journals and conferences. Dr. Sen is on the editorial board of IET Image Processing and Springer's Circuits, Systems and Signal Processing. He has received a young scientist award from The Institution of Engineers (India), a Qualcomm Innovation Fellowship, an ERCIM Alain Bensoussan Fellowship, a Ministry of Manpower (Singapore) Research Fellowship and a couple of best paper awards from IET.

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: Zhisheng Yan (zyan@gsu.edu), Yao Liu (yaoliu@binghamton.edu), Wenming Cao (wmcao@szu.edu.cn), and Phoenix Fang (dofang@calpoly.edu). 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

Chair: Jun Wu, Fudan University, China
Steering Committee Chair: Joel J. P. C. Rodrigues, Federal University of Piauí (UFPI), Brazil
Vice Chair – America: Shaoen Wu, Illinois State University, USA
Vice Chair – Asia: Liang Zhou, Nanjing University of Post and Telecommunications, China
Vice Chair – Europe: Abderrahim Benslimane, University of Avignon, France
Letters & Member Communications: Qing Yang, University of North Texas, USA
Secretary: Han Hu, Beijing Institute of Technology, China
Standard Liaison: Guosen Yue, Huawei, USA

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.