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

MMTC Communications – Review

Vol. 10, No. 6, December 2019

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the December 2019 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises three reviews that cover multiple facets of multimedia communication research including trust assessment in vehicular social network, video-based depth prediction, and weakly labeled sound event detection. These reviews are briefly introduced below.

The first paper, published in IEEE Transactions on Multimedia and edited by Dr. Jingbo Xiong, proposed a solution to trust assessment in vehicular networks, based on social computing techniques. The proposed solution could also be applied to other types of mobile social networks.

The second paper is published in IEEE Transactions on Multimedia and edited by Dr. Carl James Debono. The authors propose a solution they call Bayesian DeNet to predict the depth in monocular video. They use a deep CNN that in parallel generates a depth map and a corresponding uncertainty map.

The third paper, published in IEEE/ACM Transactions on Audio, Speech, and Language Processing and edited by Dr. Ye Liu, introduced multiple instance learning (MIL) into the study of sound event detection. Rather than annotating every short audio frame, long audio sequence with MIL could be considered a set of short audio instances and only one label is needed. As such, it significantly reduces the burden of data annotation.

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

Assessing Vehicle Trustworthiness based on Social Computing

A short review for "Trust Assessment in Vehicular Social Network Based on Three-Valued Subjective Logic"

Edited by Jinbo Xiong

T. Cheng. G. Liu, Q. Yang, and J. Sun, " Trust Assessment in Vehicular Social Network Based on Three-Valued Subjective Logic," IEEE Transactions on Multimedia, vol. 21, vo. 3, Mar. 2019.

With the rapid development of connected and autonomous vehicle technologies, it is expected that massive amount of sensor data will be exchanged among autonomous vehicles. Among the various types of applications, cooperative perception [1], is believed to have a great potential to change self-driving cars and future mobility systems. With data shared from other vehicles, a vehicle's sensing range and field of view can be significantly extended, offering a better perception about the surrounding environments [2].

Due to hardware and/or software defects, the trustworthiness of data generated by individual vehicles cannot be guaranteed. Untrustworthy information generated and exchanged from vehicles could harm the cooperative system, where vehicles may make wrong decisions and potentially cause traffic accidents. If the trustworthiness of data sent from other vehicles is unknown, a vehicle that receives the data must drop the data because it is too risky to trust the received data. To tackle this issue, the authors propose a trust assessment mechanism for vehicles, based on the trust relations between vehicle [3]. To this end, the event of exchanging data between vehicles is considered a type of social interactions between them. As vehicles are socially connected, the trust relations between them can be quantified and leveraged to assess the trustworthiness of data generated from autonomous vehicles.

In this paper, the authors first assume a static network among vehicles, i.e., the trust connections between vehicles do not change. With the three-valued subjective logic trust model [4], the trust relations between vehicles are model as trust opinions. Each trust opinion is consisted of four probability values, which are determined by whether they shared trustworthy (or untrustworthy) data between each other. With the vehicular social network in place, the authors make use of the OpinionWalk algorithm to infer the indirect trust between vehicles. Then, the authors release the assumption that trust relations between vehicles must be static, i.e., the connections between vehicles can be dynamic. Because the connections between vehicles frequently change, it poses significant challenges to the trust inference algorithm, as well as introducing large overhead on the trust assessment module. To address this issue, the authors propose to divide a vehicular social network into several communities, and apply trust assessment on vehicles within a community, and those between different communities, if necessary. Because vehicles tend to be close to each other within a certain geographic area, the resulting communities are likely to be constant and small-scale. To evaluate the proposed solution, the authors design and develop a vehicular social network simulator, and employ state-of-the-art algorithms, e.g., OpinionWalk [5], to assess the trustworthiness of vehicles.

One major contribution of this paper is to adopt trust model established in social network to qualify the trustworthiness of vehicles, regarding to sharing trustworthy data to others. This is an innovative idea as trust relations exist among vehicle, if and only if they are exchanging data. On the other hand, such interactions can be recorded and used to measure the trust relations. After a vehicle consumes the receive information, the trustworthiness of the data can be determined, which provides a means to model the trust relation between the information producers and consumers. With the state-of-the-art trust model, three-valued subjective logic, the authors essentially build a trust social network among vehicles. The social connections between vehicles occur only when information is exchange between them. Vehicles could communicate to each other, if they are within each other's communication range; however, they might not exchange data between each other necessarily. For example, a vehicle may be selfish and is not willing to share data to others. As such, a trust vehicular social network is a sub-graph of a wireless vehicular network.

The second contribution of this paper is to design the OpinionWalk algorithm to address the massive trust assessment problem. In this algorithm, the social connections between vehicles are treated as edges in a graph, which are represented in an opinion matrix. Then, a set of trust operations are introduced to compute vehicles' trustworthiness, particularly, for those that are not directly connected (or not directly interact with each other). The algorithm starts from a vehicle and searches its social network in an iterative manner to infer the trustworthiness of other vehicles.

The third contribution made by this paper is to deal with dynamic trust assessment in vehicular networks. Because vehicles keep moving on road, not only the network topology but also the trust relations in a vehicular social network will frequently change. To address this issue, the authors propose to divide a vehicular social network into several loosely connected social communities. As no prior information about the number and size of communities are known in a vehicular social network, the authors select the asynchronous label propagation algorithm, a semi-supervised method, to detect communities in a vehicular social network in a linear time. For vehicles from different communities, they will update their social trust network if they encounter with each other on road, via exchanging their trust opinion matrices. As such, the trust assessment of vehicle from different communities can be accomplished. This is reasonable as intra-community trust assessment is only necessary when vehicles from different communities meet and need to evaluate the trustworthiness of each other. For those that never leave a community, it is not necessary to employ intra-community trust assessment.

The authors finally validate the proposed solution and evaluate its performance in a simulator where vehicles movement follow the random way point model and the trust relation among vehicles are constructed when vehicles encounter with and exchange data between each other. Simulation results indicate that the proposed solution offers a more accurate trust assessment and a quicker assessing time.

References:

[1] Q. Chen, S. Tang, Q. Yang and S. Fu, "Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds," 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 2019, pp. 514-524.

[2] Chen, Q., 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 ACM/IEEE Symposium on Edge Computing, pp. 88-100, 2019.

[3] Yang, Q., and H. Wang. "Towards Trustworthy Vehicular Social Network." IEEE Communication Magazine 53, no. 8 (2015): 42-47.

[4] Liu, G., Yang, Q., Wang, H., Lin, X. and Wittie, M.P., Assessment of multi-hop interpersonal trust in social networks by three-valued subjective logic. In IEEE INFOCOM 2014, pp. 1698-1706.

[5] Liu, G., Chen, Q., Yang, Q., Zhu, B., Wang, H. and Wang, W. Opinionwalk: An efficient solution to massive trust assessment in online social networks. In IEEE INFOCOM 2017, pp. 1-9.

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 the Ph.D. degree in Computer System Architecture from Xidian University, China, in 2013. His research interests include cloud data security, mobile media management, privacy protection, and Internet of Things. He has published papers in prestigious journals such as 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 is a member of IEEE.

A Video-based Depth Prediction Method using a Single Camera

A short review for "Bayesian DeNet: Monocular Depth Prediction and Frame-wise Fusion with Synchronized Uncertainty"

Edited by Carl James Debono

X. Yang, Y. Guo, H. Lou, C. Liao and K.-T. Cheng, "Bayesian DeNet: Monocular Depth Prediction and Frame-wise Fusion with Synchronized Uncertainty," IEEE Transactions on Multimedia, vol. 21, no. 11, pp. 2701 – 2713, November 2019.

Depth information provides a measure of the distance of an object from the camera. The prediction of depth information from a single camera is an important task in computer vision as it can be applied in a number of applications. These applications include, but are not limited to, augmented reality (AR), 3D object reconstruction, and 3D modeling systems. Most of the techniques found in literature use structure-from-motion (SfM) [1] or simultaneous localization and mapping (SLAM) [2] to obtain an estimate of the depth information. These methods use motion to estimate the depth by using triangulation of the same points in a pair of consecutive frames. Instead of motion, other techniques use implicit correlations in visual cues within an image and depth [3, 4].

More advanced techniques exist and these tend to use handcrafted features together with some probabilistic model to predict the depth maps from single views, such as the work in [5]. However, with the latest advances in deep learning, convolutional neural networks (CNNs) have been successfully applied with considerable improvements in the accuracy of the predicted depth value when considering their use on benchmark datasets.

The application of CNNs in depth prediction systems only finds limit work in single images and efficient techniques to extend these methods to video sequences are still a challenging task. The model uncertainty is a major challenge in such systems as the accuracy of the prediction depends heavily on the model that is used to make the decisions on the depth values. Another challenge relates to the depth hypotheses that are in reference to the camera coordinates and need to be projected on a common system for hypotheses fusion.

The authors of the original paper propose a solution they call Bayesian DeNet to predict the depth in monocular video. They use a deep CNN that in parallel generates a depth map and a corresponding uncertainty map. The pixels in the uncertainty map represent the error distribution of the depth value found in the depth map. The camera poses between frames are found using the ORB feature tracking algorithm [6]. Using this tracking scheme, the depth estimates from the previous frame are projected onto the current camera coordinates to be used in the fusion process. A Bayesian inference technique is applied that assumes that the depth estimates in different frames are independent of each other. The posterior depth distribution at each point in the frame is then found once all the estimates are determined. This gives the fused depth that is represented by the maximum of the posterior distribution. The shape of this distribution also provides the information on the uncertainty of the estimates.

The authors report that experiments on the datasets found in [7], [8] and [9] showed little increase in computation time while providing very good accuracies. Moreover, they report that the uncertainty map can be used to identify outliers in the depth estimates which can now be removed, hence decreasing errors in the depth maps. Finally, they show that the solution can be integrated with SLAM systems improving their overall performance.

The work in this field is still limited even though it has important applications. The solution needs to be extended to typical implementations in the field where multi-view cameras exist with varying levels of scene overlap between them. Moreover, better solutions that require less computation times are necessary to be able to meet higher resolutions and higher frames per second demanded by modern ultra-high definition systems. In addition, solutions that can be embedded on small devices also demand low computations due to the limited resources. Testing in real environments is also important where applications such as obstacle avoidance in robotic platforms are needed.

References:

[1] R. Szeliski, "Structure from motion," in Computer Vision. London, U.K.: Springer-Verlag, 2011, pp. 303–334.

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J.J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[3] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[4] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5162–5170.

[5] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, May 2009.

[6] R. Mur-Artal and J. D. Tardos, "ORB SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[7] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 746–760.

[8] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proceedings of the International Conference on Intelligent Robot Systems*, Oct. 2012, pp. 573–580.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal on Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

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.

Multiple Instance Learning with Auto-Pooling for Acoustic Event Detection

A short review for "Adaptive Pooling Operators for Weakly Labeled Sound Event Detection"

Edited by Ye Liu

B. McFee, J. Salamon and J.P. Bello, "Adaptive Pooling Operators for Weakly Labeled Sound Event Detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 11, November 2018.

Sound event detection and classification play an important role in many emerging applications. For example, in environmental noise monitoring application [1], the dynamic noise map would present not only the sound pressure levels among a region, but also the corresponding noise sources. It is valuable for the analysis and mitigation of noise pollution towards green and smart city. Artificial intelligent sound recognition system for animal vocalization classification would greatly mitigate the burden of animal experts, who rely on their own auditory sense for the detection and classification of animal species in bioacoustics species study [2]. Intelligent sound classification also has great potential to improve the cyber security capability in smart home and autonomous vehicle applications to defense the malicious voice commands [3], as voice is becoming a popular interface of human-machine interaction in the age of Internet of Things.

Sound event detection is typically treated as a supervised machine learning problem. All the short clips of the audio traffic are annotated as training data. Mel-frequency cepstral coefficients (MFCC) is chosen for audio feature extraction. Then, classic machine learning algorithms such as support vector machine (SVM) and K-nearest neighbors (KNN) are applied as the classifier. For example, wireless acoustic sensor networks were deployed in the DYNAMAP project [4] to monitor traffic noise pollution in the urban and suburban environments. The unknown noise signal was gathered and divided into many short frames. Then, real-time classification was performed on acoustic sensor node to differentiate the road traffic noise and anomalous noise events. Many supervised machine learning algorithms were evaluated in terms of classification performance and computational cost. More general situation was considered in [5], in which classifier was designed to identify environmental noise from car horn children playing, street music and so on.
However, the performance of current sound event detection systems based on supervised machine learning algorithms heavily rely on the labeled training dataset. It is desired to use strongly labeled data to train the sound event detection model, but data annotation jobs are both labor-intensive and cost-intensive, which is a big hurdle to the development and practical application of sound event detection systems. Thus, it is challenging to build a high-quality sound event detection system under the condition of weakly labeled dataset.

To address above problem, this paper introduced the multiple instance learning (MIL) into the study of sound event detection. Rather than annotating every short audio frame, the long audio sequence with MIL could be considered as a set of short audio instances and only one label is needed, so that significantly reducing the burden of data annotation. Now, the key research problem is transferred to design pooling operator for aggregating the dynamic predictions in a duration to form static identifications.

Six pooling operators are presented in this paper with two standard pooling and four proposed pooling operators. They are max-pooling, mean-pooling, soft-max polling [6], auto-pooling, constrained auto-pooling and regularized auto-pooling, respectively. The drawback of standard max-pooling operator for sound event detection is that the largest instance will dominate the result. As a result, the accuracy of sound detection is sensitive to training setup. The standard mean-pooling operator seems more stable since all the sound instances are equally weighted, while the performance is also mediocre. The proposed soft-max pooling operator inherites the advantages of both max-pooling and mean-pooling operator by weighted averaging all the instances, so that each

of them contributes the label likelihood to a proportionate degree.
Unfortunately, soft-max pooling operator also has its limitation as the soft-max weight bounds will converge to the uniform weight when the bag size becomes big. The auto-pooling operator is thus proposed to address this problem. By adjusing a scalar parameter, the pooling operator is able to adaptively transform between the unweighted mean, soft-max, and max operator. Based on the auto-pooling operator, two addition improved approaches, constrained auto-pooling and regularized auto-pooling, are further proposed to deal with the specific situations like the active duration requirement and individual instance penalty. Finally, multi-label learning is discussed to illustrate the generality of proposed adaptive polling operators for weakly labeled applications.

The extensive experiments were conducted with three different types of dataset to evaluate the performance of sound detection in noise monitoring application, autonomous driving and musical instrument scenario. Many interesting finding are observed. For instance, the general auto-pooling method has a better performance for the static sound detection, but it is overfit in dynamic sound detection. The characteristics of the dataset and the duration of sound event affect the accuracy of min, max and soft-max operators. The regularized auto-pool operator is the most robust one in all conditions.

In summary, this work introduced the multiple instance learning to the study of sound event detection under weakly labeled condition. The proposed adaptive pooling operators are very promising to a variety of smart applications.

References:

[10] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," in Commun. ACM, vol. 62, no. 2, pp. 68–77, Jan. 2019.

[11] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling. "Towards the automatic classification of avian flight calls for bioacoustic monitoring," in PloS one, vol. 11, no 11, 2016.

[12] M. Zhou, Z. Qin, X. Lin, S. Hu, Q. Wang and K. Ren, "Hidden Voice Commands: Attacks and Defenses on the VCS of Autonomous Driving Cars," in IEEE Wireless Communications, vol. 26, no. 5, pp. 128-133, Oct. 2019

[13] J. Socoró, F. Alías, and R. Alsina-Pagès, "An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments," in Sensors, 2017.

[14] Y. Alsouda, S. Pllana, and A. Kurti, "IoT-based Urban Noise Identification Using Machine Learning: Performance of SVM, KNN, Bagging, and Random Forest," in Proceedings of the International Conference on Omni-Layer Intelligent Systems, ACM, pp. 62-67, 2019.

[15] J. Salamon, B. McFee, P. Li, and J. P. Bello, "DCASE 2017 submission: Multiple instance learning for sound event detection," in DCASE2017 Challenge, Tech. Rep. 2017.

Ye Liu, Ph.D, is a researcher with NAU-Lincoln Joint Research Center of Intelligent Engineering, Nanjing Agricultural University. He received M.S. and Ph.D. degrees from Southeast University, Nanjing, China in 2013 and 2018, respectively. He was a visiting scholar at Montana State University, Bozeman, USA from October 2014 to October 2015. He was a visiting PhD student from February 2017 to January 2018 in the Networked Embedded Systems Group at RISE SICS (Swedish Institute of Computer Science). His research interests include wireless sensor networks, energy harvesting systems and mobile crowdsensing. He has published papers in prestigious journals such as IEEE Communications Magazine, IEEE Internet of Things, and ACM Transactions on Embedded Computing System. He was awarded the 1st place of EWSN Dependability Competition in 2019.

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 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.